

# Internal transcribed spacer primer evaluation for vascular plant metabarcoding

Andreas Kolter<sup>1</sup>, Birgit Gemeinholzer<sup>1</sup>

<sup>1</sup> University of Kassel, Department of Botany, Heinrich-Plett-Strasse 40, D-34132 Kassel, Germany

Corresponding author: Andreas Kolter ([Andreas.Kolter@ruhr-uni-bochum.de](mailto:Andreas.Kolter@ruhr-uni-bochum.de))

Academic editor: Hugo de Boer | Received 2 May 2021 | Accepted 9 August 2021 | Published 2 September 2021

## Abstract

The unprecedented ongoing biodiversity decline necessitates scalable means of monitoring in order to fully understand the underlying causes. DNA metabarcoding has the potential to provide a powerful tool for accurate and rapid biodiversity monitoring. Unfortunately, in many cases, a lack of universal standards undermines the widespread application of metabarcoding. One of the most important considerations in metabarcoding of plants, aside from selecting a potent barcode marker, is primer choice. Our study evaluates published ITS primers *in silico* and *in vitro*, through mock communities and presents newly designed primers. We were able to show that a large proportion of previously available ITS primers have unfavourable attributes. Our combined results support the recommendation of the introduced primers ITS-3p62pIF1 and ITS-4unR1 as the best current universal plant specific ITS2 primer combination. We also found that PCR optimisation, such as the addition of 5% DMSO, is essential to obtain meaningful results in ITS2 metabarcoding. Finally, we conclude that continuous quality assurance is indispensable for reliable metabarcoding results.

## Key Words

barcoding, DMSO, internal transcribed spacer, Spermatophyta, mock community, PCR, Tracheophyta

## Introduction

Globally, one million species are threatened by extinction in the near future and 68% of monitored populations are declining (IPBES 2019; Grooten et al. 2020). An estimated two out of five plant species are threatened with extinction (Antonelli et al. 2020). Accurate monitoring is vital to understand and alleviate the driving forces behind the unprecedented biodiversity decline (IPBES 2019; Grooten et al. 2020). The term metabarcoding, which describes the analysis of complex DNA samples with the aim of taxonomic identification, has the potential to provide a scale and accuracy in biodiversity surveys that was previously unattainable for many taxonomic groups (Deiner et al. 2017; Ruppert et al. 2019). However, the increase in technical complexity (Piper et al. 2019), compared to most other monitoring methods (Marsh and Trenham 2008; Prosekov et al. 2020), also implies a higher susceptibility to errors and therefore requires stringent

quality control (Deiner et al. 2017; Ruppert et al. 2019; Thaling et al. 2020). The objective validation of metabarcoding methods can most efficiently be implemented through mock communities, since the composition of environmental DNA (eDNA) samples is unknown (Bjørnsgaard et al. 2017; Elbrecht and Leese 2017; Smith et al. 2017; Zhang et al. 2018; Braukmann et al. 2019; Thaling et al. 2020). eDNA studies can be performed either on an amplicon basis (metabarcoding) or on a genome basis (metagenomics). Although genomic methods bypass most PCR biases (Porter and Hajibabaei 2018; Piper et al. 2019), their efficiency currently is not on par with metabarcoding (Braukmann et al. 2019; Ruppert et al. 2019). In metabarcoding studies, the internal transcribed spacer 2 (ITS2) is widely used because it has a high success rate in species-level identification across the plant kingdom (see Kolter and Gemeinholzer 2020 for a detailed discussion of different plant markers). ITS2 also has one of the largest number of reference sequences in public DNA

sequence libraries amongst the most common plant barcode markers (Kolter and Gemeinholzer 2020). Universal criticism, based on the multi-copy nature of ITS2, can be countered by the fact that Song et al. (2012) discovered that 97% of all ITS2 variants in their analysis could only be found within a single species. Song et al. (2012) furthermore reported that intra-genomic distances between variants are smaller than intra-specific or inter-specific distances. Therefore, ITS2 remains to be an important tool in metabarcoding, phylogeny and many other applications (Kay et al. 2006; Feliner and Rosselló 2007; Cheng et al. 2016; Alanagreh et al. 2017; Liu et al. 2019).

Aside from marker choice, primer choice has repeatedly been identified as one of the key factors to facilitate accurate recovery of taxa in a sample (Krehenwinkel et al. 2017; Elbrecht et al. 2019; Hajibabaei et al. 2019; Kelly et al. 2019; Piñol et al. 2019; Li et al. 2020). However, despite the availability of multiple ITS primer sets (White et al. 1990; Gu et al. 2013; Cheng et al. 2016; Moorhouse-Gann et al. 2018; Tremblay et al. 2019), we have identified persistent problems with the amplification success of ITS in multiple recent large-scale barcoding studies (Braukmann et al. 2017; Gill et al. 2019; Jones et al. 2021). The lack of a side-by-side evaluation of current ITS primers makes it impossible to identify and solve the underlying issues of low ITS amplification success rates, compared to other barcode markers.

Our study evaluates ITS primers, based on an *in silico* and *in vitro* analysis. The *in vitro* analysis, performed by using two mock communities, aims to compare the uniformity of amplification achieved with different primers and identify primer-specific amplification biases. The *in silico* analysis identified mismatches of common ITS primers in Spermatophyta (Cycadopsida, Gnetopsida, Pinopsida, Liliopsida and Magnoliopsida) and led to the design of five new ITS primers with improved universality. Our primer design was focused towards Spermatophyta, as this plant taxon is well represented in public sequence repositories compared to other plant taxa. However, we also reported mismatches in Bryophyta, Fungi, Polypodiopsida and Lycopodiopsida if adequate taxonomic representation were available. Furthermore, due to the high guanine-cytosine (GC) content in a substantial number of ITS2 sequences, we investigated the impact of dimethyl sulfoxide (DMSO) on mock community species retrieval success.

## Material and methods

### De novo primer design

Spermatophyta sequences containing the ITS region, used as a template to generate primers, were downloaded from GenBank in April 2018 as described in Kolter and Gemeinholzer (2020). Degenerate consensus ITS sequences on family level of the nrDNA LSU (large subunit of the nuclear ribosomal DNA) and nrDNA SSU (small

subunit of the nuclear ribosomal DNA) regions were used to identify conserved flanking regions and, subsequently, suitable primer locations in a single step.

To screen for potential primer sequences in the 5.8S nrDNA region, a consensus sequence of all Spermatophyta plant sequences was established (Suppl. material 1: Suppl. file 1). All nucleotides with more than 2% abundance at a specific position were taken into consideration and represented by the respective IUPAC code. Following this, all 24-mers, extracted from the aforementioned single consensus sequence with a maximum of eight IUPAC ambiguity codes and ending in a C or G, were identified as primer candidates. Primer candidates with hairpin structures with an average melting temperature above 50 °C and those with a GC content below 40% or above 80% were filtered out. Primer candidates forming self-dimers or hetero-dimers with any reverse primer (created in this work), with a Gibbs free energy ( $\Delta G$ ) higher than one fourth of the maximum Gibbs free energy were discarded or modified. The remaining primer candidates were aligned with 5.8S nrDNA consensus sequences on family level and mismatches were manually resolved by adding a degenerate nucleotide code, if possible. Some mismatches were specific for certain plant families and, therefore, could not be resolved without overly inflating the overall primer degeneracy (Suppl. material 1: Suppl. file 2). The three forward primers have been named in accordance with White et al. (1990) followed by the position (p) within the 5.8S nrDNA region, the specificity (pl = plant, un = universal), the orientation (F = forward, R = reverse) and a revision number (e.g. ITS-3p53plF1). Previously published primers located in the 5.8S nrDNA region played no role in primer positioning or design.

### *In silico* primer evaluation

Primer statistics were calculated using the online tool OligoAnalyzer 3.1 (Owczarzy et al. 2008) with the following settings: monovalent cations ( $K^+$ ) 50 mM, divalent cations ( $Mg_2^{+}$ ) 2.5 mM, dNTP 0.05 mM and primer 0.2  $\mu M$ . Primer melting temperatures were calculated using the method of Allawi and SantaLucia (1997).

We selected 22 frequently used ITS primers from literature to be analysed alongside the five newly designed primers (Table 1). Primers were validated with Tracheophyta sequences downloaded from GenBank in September 2020 and subsequently processed as described in Kolter and Gemeinholzer (2020). Additional sequence alignment filter steps on family level included the removal of columns with more than 95% gap characters and alignment columns that were supported by less than three species. The sequence coverage of the LSU and SSU region varied from 22,574 to 63,024 (SSU) and from 25,845 to 90,319 (LSU) due to the fact that regions located upstream of the most common ITS primers were less covered in GenBank (Appendix 1). Primer BEL-3, designed by Chiou et al. (2007) and referred to as S3R by

**Table 1.** Overview of primer sequences used in this study.

Primer name	Orientation	nrDNA primer position	Distance to ITS region from primer 3' [bp]	Primer sequence (5'→3')	Publication	<i>in silico</i> / <i>in vitro</i> evaluation
ITS1	forward	SSU	11	TCCGTAGGTGAACCTGCGG	White et al. (1990)	✓ / X
ITS5	forward	SSU	32	GGAAGTAAAAGTCGTAACAAGG	White et al. (1990)	✓ / X
ITS-A	forward	SSU	32	GGAAGGAGAAGTCGTAACAAGG	Blattner (1999)	✓ / X
ITS-u1	forward	SSU	32	GGAAGKARAAGTCGTAACAAGG	Cheng et al. (2016)	✓ / X
ITS-p5	forward	SSU	46	CCTTATCAYTTAGAGGAAGGAG	Cheng et al. (2016)	✓ / X
ITS2	reverse	5.8S	30	GCTGCGTTCTTCATCGATGC	White et al. (1990)	✓ / X
ITS-C	reverse	5.8S	55	GCAATTCACACCAAGTATCGC	Blattner (1999)	✓ / X
ITS-u2	reverse	5.8S	86	GCGTTCAAAGAYTCGATGRITC	Cheng et al. (2016)	✓ / X
ITS-p2	reverse	5.8S	5	GCCRAGATATCCGTTGCCGAG	Cheng et al. (2016)	✓ / X
ITS-2plR1	reverse	5.8S	3	GCCDAGATATCCRTTGYCRRGAG	this work	✓ / X
ITS3	forward	5.8S	110	GCATCGATGAAGAACGCAGC	White et al. (1990)	✓ / ✓
ITS-D	forward	5.8S	136	CTCTCGGCAACGATATCTCG	Blattner (1999)	✓ / X
ITS-S2F	forward	5.8S	87	ATGCGATACTTGGTGTGAAT	Gu et al. (2013)	✓ / ✓
ITS-u3	forward	5.8S	110	CAWCGATGAAGAACGYAGC	Cheng et al. (2016)	✓ / ✓
ITS-p3	forward	5.8S	141	YGACTCTCGGCAACGGATA	Cheng et al. (2016)	✓ / ✓
UniPlantF	forward	5.8S	75	TGTGAATTGCARRATYCMG	Moorhouse-Gann et al. (2018)	✓ / ✓
ITS-3p34unF1	forward	5.8S	104	CGATGAAGAAAGYAGYRAAMTG	this work	✓ / X
ITS-3p53plF1	forward	5.8S	84	AMTGCGAYACBTRGTGTGAATTGC	this work	✓ / X
ITS-3p62plF1	forward	5.8S	78	ACBTRGTGTGAATTGCAGRATC	this work	✓ / ✓
58SPL	forward	5.8S	42	TTTGAACGCAAGTTGCGCC	M.-J. Côté, published: Tremblay et al. (2019)	✓ / X
ITS4	reverse	LSU	40	TCCTCCGCTTATTGATATGC	White et al. (1990)	✓ / ✓
ITS-B	reverse	LSU	41	CTTTTCTCCGCTTATTGATATG	Blattner (1999)	✓ / X
BEL-3	reverse	LSU	144	GACGCTTCTCCAGACTACAAT	Chiou et al. (2007)	X / ✓
ITS-u4	reverse	LSU	49	RGTTTCTTTTCTCCGCTTA	Cheng et al. (2016)	✓ / ✓
ITS-p4	reverse	LSU	35	CCGCTIAKTGATATGCTTAAA	Cheng et al. (2016)	✓ / ✓
UniPlantR	reverse	LSU	2	CCCGHYTGAYYTGRGGTDCD	Moorhouse-Gann et al. (2018)	✓ / ✓
ITS-4unR1	reverse	LSU	40	TCCTCCGCTTATTKATATGC	this work	✓ / ✓

Chen et al. (2010), was not evaluated *in silico* as its distance to the ITS2 region resulted in very poor sequence availability. To evaluate the specificity towards Fungi and Bryophyta, sequences were extracted from Cheng et al. (2016). The specificity analyses of Polypodiopsida and Lycopodiopsida has been limited to the 5.8S nrDNA region (for a detailed sequence list with taxonomic information, see Suppl. material 1: Suppl. file 1).

Primers were compared to the DNA sequences using the R packages ShortRead and Biostrings (Morgan et al. 2009; Pagès et al. 2020). To give each species the same weight in primer evaluation, the sum of mismatches per sequence has been divided by the number of sequences of that respective species. Adding these numbers up on a family level and dividing them by the total number of species per respective family, resulted in a mismatch score, given in percentages (Suppl. material 1: Suppl. file 3). Only mismatch scores above 30%, per primer position, were reported (for an unfiltered list, see Suppl. material 1: Suppl. file 2). Figures were created using ggplot (Wickham 2016). Higher taxonomic names (above the rank of family) have been retrieved by the R package rgbif from the Global Biodiversity Information Facility (GBIF) backbone taxonomy and are meant to be descriptive only (Chamberlain and Boettiger 2017).

Mock community design

We extracted DNA from 58 herbarium specimen by the use of silica-coated ferric beads and the tissue protocol by Sellers et al. (2018). The two mock communities (mix 1, mix 2) were constructed by considering: 1) DNA concen-

tration, 2) individual amplification success, 3) taxonomic diversity and 4) inclusion of samples with high GC content. We did not use known mismatches as a criterion for inclusion or exclusion. Non-Spermatophyta species were added to investigate primer specificity. Mix 1 (Appendix 4) was created from 23 different Spermatophyta plus four Lycopodiopsida species (*Selaginella kraussiana*, *Selaginella denticulata*, *Huperzia carinata* and *Equisetum arvense*) and two fungal species (*Aspergillus chevalieri* and *Pseudogymnoasus pannorum*). Mix 2 (Appendix 5) was created from 20 different Spermatophyta plant species plus two fungal species (*Talaromyces wortmannii* and *Aureobasidium pullulans*). The mixtures contained equal amounts of DNA from each species (1 ng), as determined by Qubit v.4.0 (Invitrogen, dsDNA HS Assay Kit Q32854). This resulted in two mock communities with a taxonomic spread of 16 Spermatophyta plant orders and a difference in GC content of ~25% (Appendices 5 and 6). This intentionally resulted in mock communities with a wide variety of ITS copy numbers per species.

Sequencing primer design

Primers contained a part of the Illumina TruSeq read primer in addition to the target primer to act as a linker between PCR number one (target amplification) and PCR number two (Illumina indexed adapter being added), which results in the following forward primer: 5'-CAGACGTGTGCTCTTCCGATCT [optional spacer] [target primer]-3' (reverse: 5'-CTACACGACGCTCTTCGATCT [optional spacer] [target primer]-3'). A spacer that is non-complementary to the target sequence was

added to: 1) prevent more than three identical consecutive nucleotides, 2) stop the TruSeq sequence from interfering with primer binding if it showed the potential to be partially complementary to the target sequence and, 3) to act as a mini-barcode with a length of 3 bp to facilitate pooling of samples with otherwise identical primers after the first round of PCR (termed internal index by Glenn et al. (2019)). The second round of PCR targeted the primer linker added before, which also acts as a part of the read primer in the upcoming sequencing reaction and added the barcoded Illumina indexes (i7 and i5) and the p5 and p7 sequencing adapters (Appendix 3).

### PCR setup

PCRs were conducted in a 12.5 µl reaction mix containing: 3.125 µl Trehalose (20%), 1.25 µl reaction buffer (10×), 0.625 µl MgCl<sub>2</sub> (50 mM), 1.25 µl DMSO (50%), 0.3 µl bovine serum albumin (BSA) (0.01 mg/ml), 0.25 µl each forward and reverse primer (5 µM), 0.3125 mM dNTP (2 mM), 0.06 µl Platinum Taq (Invitrogen) polymerase (5 U/µl), 1 µl DNA template and 4.0775 µl ddH<sub>2</sub>O (modified from Fazekas et al. (2012)). PCR cycling conditions were 3 minutes at 95 °C for an initial denaturation, followed by 30 cycles (each: 30 sec denaturation at 95 °C, 30 sec annealing at 50 °C, 45 sec elongation at 72 °C) and followed by a final extension at 72 °C for 6 minutes. PCRs were set up in nine technical replicates and three non-template controls (Appendix 8). Five µl of three PCRs with different primer regions were pooled and subsequently treated with Exo I (Thermo Scientific, EN0582). Samples were sent to LGC Genomics GmbH (Berlin, Germany) where they were sequenced on a MiSeq (2x300bp) after an additional 12 PCR cycles (second round of PCR) to add to the indexed Illumina adapter. The second round of PCRs was performed by an initial three cycles at low annealing temperature (each: 15 sec denaturation at 96 °C, 30 sec annealing at 50 °C, 90 sec elongation at 70 °C) and followed up by nine cycles with increased annealing temperature (each: 15 sec denaturation at 96 °C, 30 sec annealing at 58 °C, 90 sec elongation at 70 °C) using MyTaq Red Mix (Bioline, BIO-25044) polymerase. The sequencing library concentrations were adjusted to approximately meet the target of 100,000 total reads per sample which translates to 50,000 paired reads and approximately 16,670 paired reads per PCR (first round of PCR). The sequencing library included all PCR control reactions which totalled approximately 25% of all samples (Suppl. material 1: Suppl. file 4).

### PCR optimisations

Due to the relatively high GC content of ITS2 amplicons, we optimised PCR conditions in a pre-trial and concluded that the additive DMSO at a concentration of 5% enables amplification of ITS2 from genomic plant templates (Suppl. material 1: Suppl. file 1). To assess the impact of DMSO on mixed PCR templates, like the mock com-

munities used in this study, we compared 0% DMSO to 5% DMSO using the ITS-3p62pIF1 + ITS-4unR1 primer combination. All other parameters were identical to the protocols mentioned earlier.

### Sequence data analysis

Sequencing data was processed with R (Suppl. material 1: Suppl. file 3) and VSEARCH (Rognes et al. 2016; R Core Team 2020). All filtering steps were applied to individual reads instead of read pairs (paired forward and reverse reads) to retain reads in which one of the sequencing directions did not produce any or not enough data. Such reads were granted permission to bypass the merge step if they had a minimum length of 150 bp after primer removal and quality control. In case both reads passed the quality filters and length requirement, but could not be merged, the longer read was retained and the corresponding paired read was discarded. The first step of filtering raw reads was performed by R and included the removal of sequences showing errors in the primer sequence or sequences that were shorter than 30 nucleotides. Subsequent filtering steps by vsearch removed all reads with any undetermined nucleotides (N) and truncated the reads after the continuously accumulated chance of an erroneously assigned nucleotide reached one (maximum expected errors). The merge step allowed only reads to pass (for exceptions, see above) which showed a minimum overlap of 20 nucleotides with a maximum of five differences and which produced a merged sequence with a minimum length of 60 nucleotides. Merged reads were deduplicated by vsearch at 100% identity in a reversible manner and identified by SINTAX (Edgar 2016) using a database by Ankenbrand et al. (2015). Obvious misidentifications (i.e. due to missing reference sequences) were manually corrected. Mock communities were analysed on a genus level.

### Sequence data derived metrics

To assess the successful detection of taxa in the mock communities, we calculated its read abundance for each taxon and primer combination. We define the read abundance for each taxon in each replicate as the proportion of reads for a given taxon relative to 1000 reads. If the median read abundance of all replicates of one taxon of a specific primer combination was above 0.1 and the taxon was detected ( $\geq$  one read) in all replicates of the respective primer combination, the taxon was classified as present (Table 3; Appendices 4 and 5). We set a minimum of more than one read in 10,000 per replicate as the detection threshold as, based on the read depth, this requires a taxon to be not represented by singletons only. We furthermore calculated a retrieved taxa score by including partial detections (taxon not detected in all replicates). The most optimal outcome would be for all taxa within one mock community to be detected in all replicates. The retrieved taxa score expresses how many of those detec-

**Table 2.** Primer statistics of this study.

Primer name	Spermatophyta families with mismatches (total no. of families tested)	Average# number of mismatches in fungi sequences	Melting temperature [°C] (min / mean / max)	GC content [%] (mean/ max)	GC terminal 3' stretch [bp]	Primer length [bp]	Max ΔG [kcal/mole]	Self-dimer ΔG [kcal/mole]	Hairpin melting temperature [°C] mean/max	Max repeats [n]*
ITS1	5 (149)	< 1 – 2	65.8	63	4	19	-39.83	-6.68	66.4	2
ITS5	113 (116)	< 1	59.3	41	2	22	-38.61	-3.61	29.5	4
ITS-A	5 (116)	1 – 2	62.3	50	2	22	-40.23	-3.61	29.5	2
ITS-u1	0 (116)	< 1	59.4 / 60.8 / 62.3	45 / 50	2	22	-39.42	-3.61	29.5	4
ITS-p5	1 (95)	> 3	57.8 / 58.5 / 59.2	43	1	22	-37.86	-6.61	44.1	3
ITS2	26 (210)	< 1	64.0	55	2	20	-40.18	-13.62	37.8	2
ITS-C	33 (210)	> 3	62.2	48	3	21	-39.29	-5.36	25.2	2
ITS-u2	71 (210)	< 1	60.5 / 62.1 / 63.7	45 / 50	1	22	-41.25	37.8	50.5	3
ITS-p2	44 (210)	> 3	65.1 / 65.8 / 66.5	60 / 62	1	21	-44.32	-7.06	42.8	2
ITS-2pIR1	8 (210)	> 3	60.5 / 65.0 / 69.9	54 / 65	1	23	-44.15	-7.15	26.7 / 56.3	3
ITS3	26 (210)	< 1	64.0	55	2	20	-40.18	-13.62	37.3	2
ITS-D	56 (210)	> 3	63.9	57	2	21	-41.93	-7.06	64.1	2
ITS-S2F	36 (210)	> 3	60.1	40	0	20	-35.68	-3.61	34.7	2
ITS-u3	22 (210)	< 1	58.2 / 59.9 / 61.8	50 / 53	2	19	-35.76	-13.74	32.7 / 47.0	2
ITS-p3	30 (210)	> 3	62.8 / 63.1 / 63.4	55 / 58	0	19	-37.80	-5.19	64.1	2
UniPlantF	8 (210)	< 1	55.4 / 58.7 / 62.5	42 / 53	4	19	-35.49	-10.76	29.9 / 65.1	4
ITS-3p34unF1	5 (210)	< 1	55.9 / 61.0 / 65.8	43 / 55	1	22	-39.94	-6.48	34.0 / 57.3	4
ITS-3p53plF1	9 (210)	< 1 – 2	62.3 / 66.1 / 69.5	47 / 54	2	24	-44.47	-8.65	48.5 / 77.4	2
ITS-3p62plF1	9 (210)	2 – 3	59.4 / 62.5 / 65.3	44 / 50	1	22	-38.83	-7.05	29.1 / 58.4	2
58SPL	11 (210)	1 – 2	64.7	53	5	19	-41.15	-10.65	65.6	3
ITS4	14 (130)	< 1	59.8	45	2	20	-38.09	-3.91	28.8	2
ITS-B	33 (123)	< 1	59.8	39	1	23	-42.38	-3.91	9.9	4
BEL-3	NA (NA)	NA	61.7	48	0	21	-36.77	-3.61	29.1	2
ITS-u4	28 (109)	< 1	59.7 / 60.0 / 60.4	42 / 45	0	20	-38.71	-3.61	17.3	4
ITS-p4	18 (140)	1	57.1 / 57.8 / 58.6	36 / 38	0	21	-38.43	-4.85	34.8 / 40.6	3
UniPlantR	33 (176)	> 3	60.3 / 65.8 / 73.0	63 / 80	3	20	-41.13	-10.71	42.9 / 56.0	3
ITS-4unR1	5 (130)	< 1	57.5 / 58.7 / 59.8	42 / 45	2	20	-37.78	-3.91	20.3	3

\*: reports the maximum number of either mononucleotide or dinucleotide repeats #: We estimated the number of mismatches, based on available sequences.

tions, in relation to the maximum number of possible detections (i.e. mix 1: 21 \* 9 replicates), were successful by a specific primer combination.

The required read depth for each primer combination to detect all but one taxon with a confidence of 95% within a mock community was calculated separately for mix 1 and mix 2 in multiple steps. First, instead of assigning an arbitrary penalty score to missing values (taxon not detected in some replicates), we limited the calculation to taxa which could be detected in all replicates in all primer combinations (Appendices 4 and 5). Second, we subsampled the reads of each replicate for each primer combination in increments of 100 reads, each 100 times. If the specific taxon was found ( $\geq$  one read) in at least 95 of the 100 repeated subsamples, it was scored as detected. For example, in one case, we subsampled the reads of replicate three out of nine (mix 1) for the primer combination ITS-3p62plF1 + ITS-4unR1 100 times at a depth of 3000 reads. In this example, *Solanum citrullifolium* was detected in 88 out of the 100 subsamples and, therefore, was scored as “not detected”. Third, the lowest number of subsampled reads required to achieve the aforementioned detection rate of 95% for each replicate and each taxon per primer combination was averaged per taxon. We finally took the 95<sup>th</sup> percentile of the aforementioned average values per primer combination to estimate the number of reads required to detect all taxa in that respective primer combination with a confidence of 95%, except one (give outliers less weight), which was allowed a lower confidence score (Table 3).

Due to the sample size of the mock communities, the 95<sup>th</sup> percentile was roughly equivalent to the average

value of the two taxa requiring the most reads. As the removal of singletons is desirable in some experimental setups, we cloned the workflow mentioned before with the exception that two reads of a specific taxon had to be detected in 95 out of 100 subsamples per primer combination (Table 3). If the number of reads of any primer combination were not sufficient to achieve 95% detection probability for some difficult to detect taxa with low read abundances, we linearly interpolated the read abundances to increase the sample size.

## Results

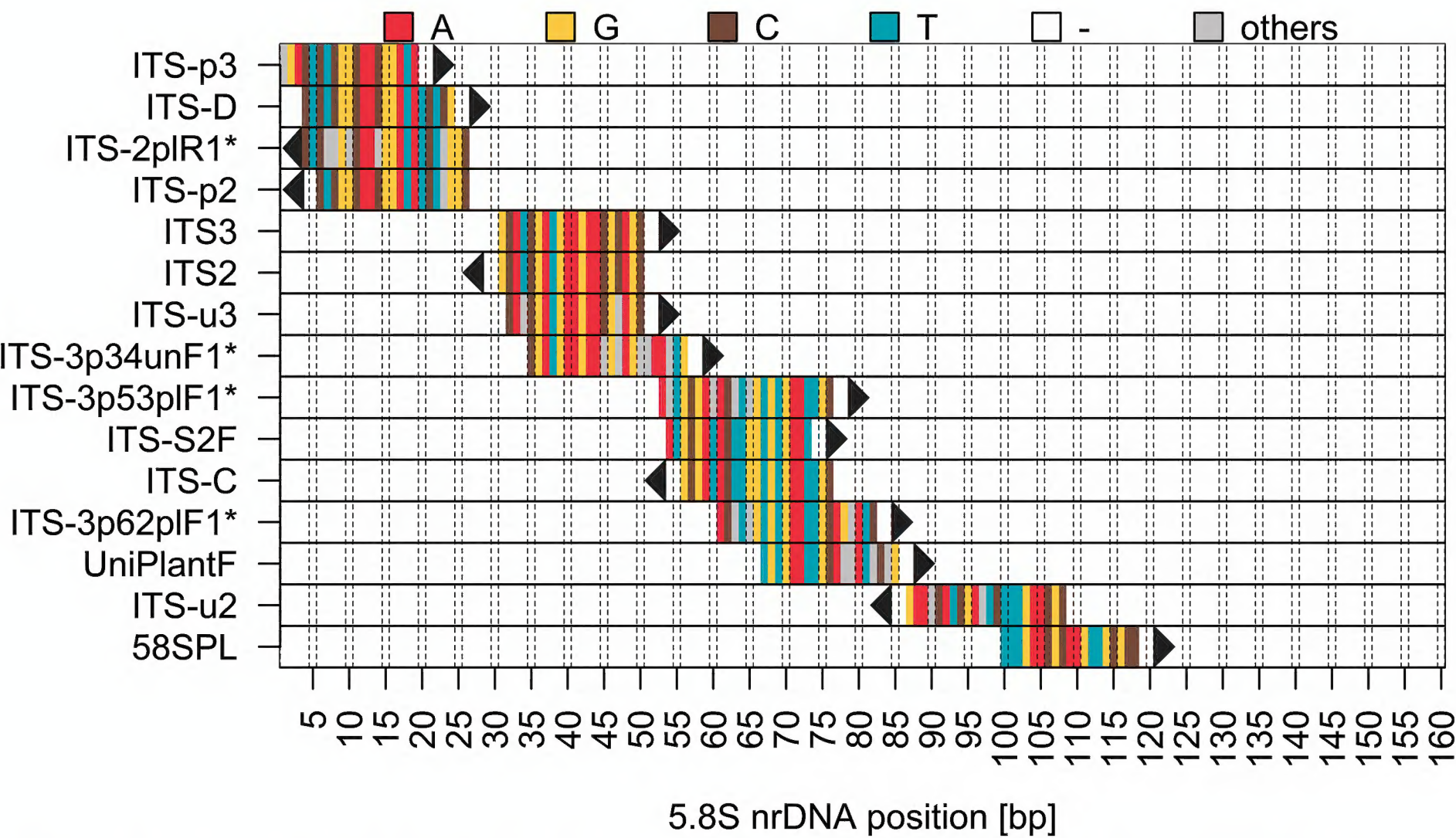
### *In silico* primer evaluation

We tested a total of 26 primers (Table 1), located in three distinct regions (SSU, 5.8S nrDNA and LSU). Primer melting temperatures of ITS primers analysed in this study ranged from 55.4 °C to 73.0 °C (Table 2). With four exceptions (ITS1, ITS-D, ITS-p3 and 58SPL), the hairpin melting temperature was below the melting temperature of the primer itself (Table 2). Some of the primer variants of primers featuring ambiguous nucleotides (i.e. UniPlantF and ITS-3p53plF1), also form hairpin structures with a melting temperature higher than their respective melting temperature (Table 2). However, our results indicate that these primer variants do not match any plant template in our database (Supp. file 5). We listed the plant families that are potentially negatively impacted by hairpin structures with a melting temperature higher than 50 °C (Suppl. material 1: Suppl. file 6).

**Table 3.** Mock community key attributes.

Primer combination		Mix 1 (n = 21)						Mix 2 (n = 18)				
Forward	Reverse	fungi reads [%]	Lycopodiopsida reads [%]	missed taxa (n)	retrieved taxa score [%]	min. required read depth (incl. singletons)	min. required read depth (excl. singletons)	fungi reads [%]	missed taxa (n)	retrieved taxa score [%]	min. required read depth (incl. singletons)	min. required read depth (excl. singletons)
ITS-3p62plF1	ITS-4unR1	< 1	5	3	92	2,000	3,300	1	3	94	1,300	2,000
UniPlantF	UniPlantR	< 1	0	7	81	5,500	9,600	< 1	5	81	2,400	3,400
UniPlantF	ITS-4unR1	41	3	6	85	3,200	5,100	47	5	86	2,400	5,600
ITS-3p62plF1	UniPlantR	< 1	0	5	88	5,300	9,400	< 1	4	83	1,900	2,700
58SPL	ITS-4unR1	10	< 1	6	81	4,800	7,600	16	5	80	1,300	2,100
ITS-u3	ITS-u4	57	13	6	90	6,600	10,800	74	4	87	6,600	9,800
ITS-p3	ITS-p4	< 1	34	6	83	2,400	4,000	< 1	5	83	1,800	2,700
ITS-S2F	BEL-3	< 1	26	7	83	6,800	11,200	< 1	3	89	5,200	7,600
ITS3	ITS4	80	6	9	75	8,800	14,300	87	6	77	9,100	14,000

Note: The top three most optimal values in each column were highlighted (bold) until three values were chosen and the next higher or lower value is different from the previous one. Lycopodiopsida read proportions were not highlighted, as it depends on the particular study whether they are negatively (non-target taxon) or positively (target taxon) evaluated. A taxon was defined as missing if the read abundance were lower than 1 in 10000 reads or if it was not represented in all technical PCR replicates. The retrieved taxa score, in contrast, includes taxa that were represented in less than all technical PCR replicates. The required read depth only considers taxa that could be fully recovered (present in all replicates) in all primer combinations. Values were rounded, but never to zero, if at least one read could be detected.



**Figure 1.** Position of previously published and newly introduced ITS primers in the 5.8S nrDNA region (5' → 3'). Primer positions may be shifted by ± 2 bp in comparison to previously published alignments due to different annotation software being used to identify the 5.8S region. Primers usually used to amplify the ITS1 region (Table 1) have been reverse-complemented to fit the 5' → 3' plus strand orientation of the figure. A black arrowhead indicates the direction of amplification of the original primer sequence (Table 1). Primers introduced in this work are marked by an asterisk.

GC primer content ranged from 36% to 80% with a 3' terminal GC stretch of zero to five (Table 2). The UniPlantR primer variant (CCCGCCTGACCTGGGGTCGC) that matches with the most (~72%) plant template sequences in our database, has a GC content of 80% (Suppl. material 1: Suppl. file 5). The primer lengths varied between 19 bp and 24 bp and resulted in a maximum ΔG between -44.32 and -35.49 kcal per mole (Table 2). The self-dimer ΔG ranged from -13.74 to -3.61 kcal per mole (Table 2). The smaller the ratio between maximum ΔG and self-dimer ΔG, the less likely it is that the primer forms troublesome

dimers. The number of mononucleotide and dinucleotide repeats ranged from two to four (Table 2). We identified four primer hotspots within the 5.8S nrDNA region (Fig. 1). Their central motifs are approximately located at the positions: 5–20 bp (ITS-D, ITS-p3, ITS-p2, ITS-2pIR1), 30–50 bp (ITS3, ITS-u3, ITS-3p34unF1), 60–75 bp (ITS-S2F, UniPlantF, ITS-3p53plF1, ITS-3p62plF1) and 100–108 bp (ITS-u2, 58SPL).

The total amplicon length of each primer combination can be calculated by adding the primer lengths, the distance to the ITS region of interest and the length of the

ITS marker (Table 1). Most ITS2 sequences have a length between 183 bp (0.5% percentile) and 271 bp (99% percentile), with a median length of 220 bp (Kolter and Gemeinholzer 2020).

### ***In silico* primer evaluation of the SSU region**

Regardless of their position, the primers located in the SSU region, flanking the ITS1 region (ITS1, ITS-A, ITS-u1, ITS-p5), have five or fewer mismatches with the exception of ITS5 (Tables 1 and 2; Suppl. material 1: Suppl. file 2 and Suppl. material 1: Suppl. file 3). However, a stable hairpin structure, with a melting temperature above the annealing temperature, is formed by a 7 bp long self-complementary stretch of the ITS1 primer (Table 2).

### ***In silico* primer evaluation of the 5.8S nrDNA region (reverse)**

The ITS-2pLR1 primer displays the lowest number of mismatches (Table 2, Suppl. material 1: Suppl. file 3) in comparison to other primers in the 5.8S region that are generally used as reverse primers for the ITS1 region. The ITS-2pLR1 primer is similar to the ITS-p2 primer, the introduction of ambiguous base pairs reduces the number of Spermatophyta families with mismatches from 44 to 8 (Table 2). The remaining eight primer mismatches can be found in the plant families Potamogetonaceae, Apiaceae, Plumbaginaceae, Thismiaceae, Siparunaceae, Melanthiaceae, Eriocaulaceae and Juncaceae (Suppl. material 1: Suppl. file 3). The primer with the next lowest number of mismatches, ITS2, with mismatches in 26 families (Table 2), features a stretch of eight self-complementary bases located near the 3' end and, therefore, bears a high risk of producing unwanted by-products.

### ***In silico* primer evaluation of the 5.8S nrDNA region (forward)**

The primers located in the 5.8S nrDNA region that are usually used to serve as forward primers to amplify the ITS2 region can be split into two major groups by the number of Spermatophyta families with mismatches (Table 2). Group one (ITS3, ITS-D, ITS-S2F, ITS-u3 and ITS-p3) does not match in a minimum of 22 families, while group two (UniPlantF, ITS-3p\* and 58SPL) has a maximum of 11 mismatched families (Table 2). Most mismatches were found in families outside of Magnoliopsida, like Orchidaceae (Suppl. material 1: Suppl. file 3). The ITS-3p34unF1 primer featured the lowest number of plant families with mismatches: Orchidaceae, Cordiaceae, Thismiaceae, Siparunaceae and Typhaceae (Suppl. material 1: Suppl. file 3).

### ***In silico* primer evaluation of the LSU region**

Primers located in the LSU region, except for UniPlantR, are overlapping each other at a distance of 35–49 bp to

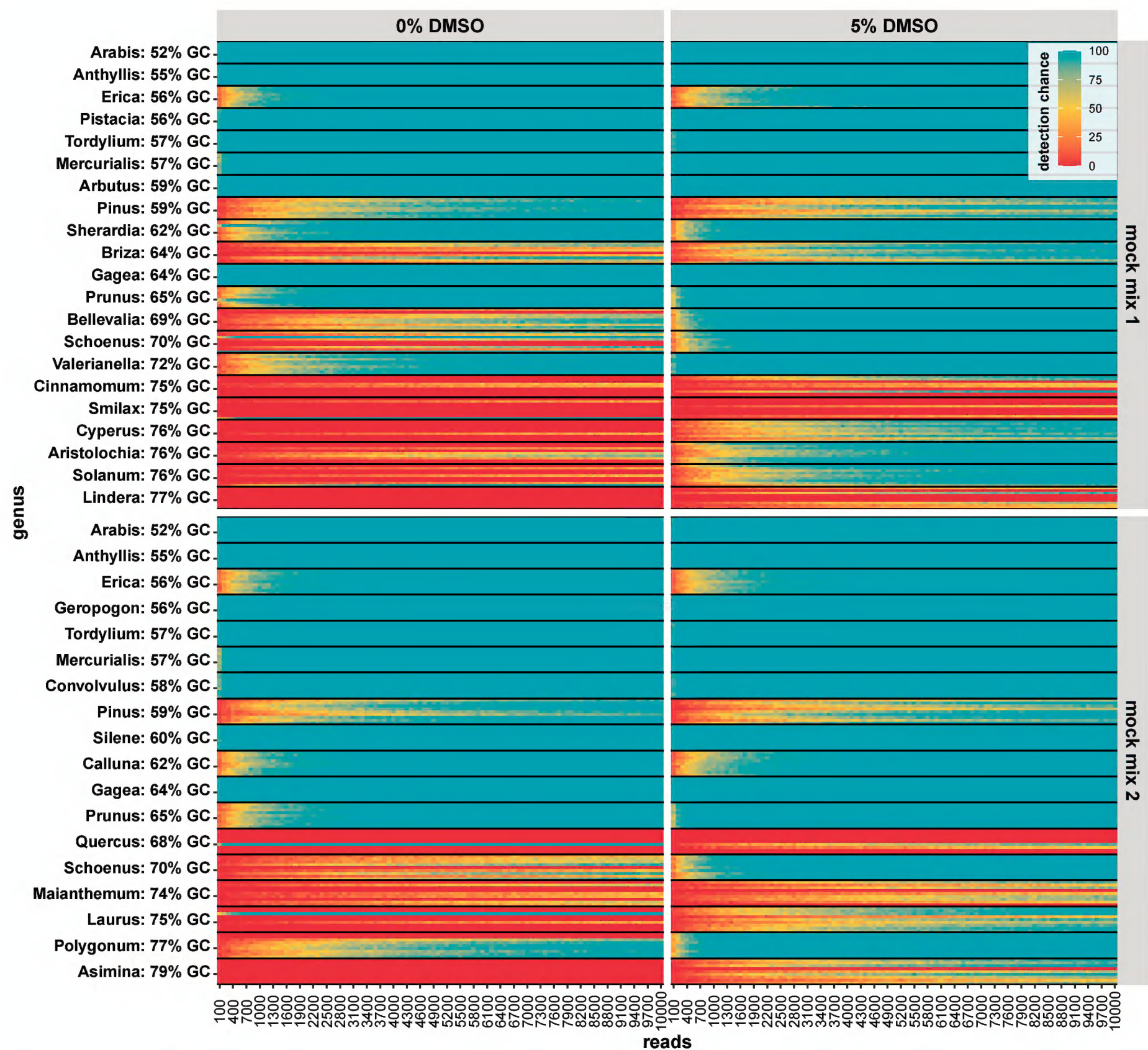
the ITS2 region (Table 1). A modification of the ITS4 primer, ITS-4unR1, displays half the amount of mismatched Spermatophyta families when compared to other primers in the LSU region (Table 2, Suppl. material 1: Suppl. file 3). The remaining mismatches in the families Gnetaceae, Araucariaceae, Cupressaceae, Pinaceae and Juncaceae can be eliminated by adding one additional ambiguous nucleotide at 11T>D. (Suppl. material 1: Suppl. file 3).

### **Mock community sequence analysis**

The sequencing yielded 3,196,249 paired raw reads (NCBI BioProject PRJNA740294). The filtering step that removed most reads on average (~30%) was to require an exact primer fit at the start of the sequence. A total of 6.5% of all reads failed to merge. As the overall read quality was very high, other filters (combined) removed, on average, less than 5% of the total reads. This results in a total of 2,128,039 merged reads to enter the analysis. There was no contamination detected in the blanks that could have affected the results. Rarefaction curves show a flat slope, except for the primer combination ITS-p3 + ITS-p4, as it yielded less reads than the other primer combinations (Suppl. material 1: Suppl. file 4). All primer combinations are represented by nine replicates, with the exception of ITS-S2F + BEL-3 which is represented by three replicates only. There is a negative Spearman correlation of -0.7 and -0.67 between the GC content and the number of reads recovered for each taxon in mix 1 and mix 2, respectively ( $p < 0.01$ ,  $df = 16$ ). This is expected, as species with a relatively high GC content were intentionally included (Appendices 4 and 5). The GC content of taxa ranged from 52% to 77% in mix 1 and from 52% to 79% in mix 2 (Appendices 4 and 5). We recovered no correlation between the number of reads for a specific taxon and the number of mismatches to its respective primer pair. The number of reads from unexpected taxa (reads from species that were not fungi and not included in the mock communities) were negligible with less than 1 in 1,000 reads.

### **Mock community primer tests**

*Philodendron angustisectum* could not be detected by any primer combination targeting the ITS2 region in mix 1 and *Sassafras albidum* could not be detected in mix 2, most probably due to DNA degradation. Both will be excluded from the following analysis. The number of missed taxa per mix ranges from 3 to 9 out of 21 for mix 1 and from 3 to 6 out of 18 for mix 2 (Table 3). The number of retrieved taxa is highest with the ITS-3p62plF1 + ITS-4unR1 primer combination and lowest with ITS3 + ITS4 (Table 3). The same pattern in reverse was observed for the required read depth for both analyses, with and without singletons (Table 3). Primer combinations (mix 1 and mix 2) that contain the reverse primer UniPlantR generally recovered very low reads of Liliopsida, especially



**Figure 2.** Impact of DMSO on mock community representation. The detection chance (colour) per genus (black lines) was tracked per replicate (horizontal coloured lines) by subsampling the reads 100 times randomly in steps of 100 from 100 to 10000 (x-axis). The detection chance was defined as the number of subsamples where the respective genus could be detected by at least one read.

*Gagea graeca*, which becomes nearly undetectable with this primer (Appendices 4 and 5).

The 58SPL primer has a similar number of mismatches compared to 3p62plF1 (Table 2). Yet, in direct comparison, we find the key metrics clearly in favour of the 3p62plF1 primer (Table 3). Taking a closer look, the taxa showing a read abundance greater than 200 in the 3p62plF1 + ITS-4unR1 primer combination are increasing in read abundance in the 58SPL + ITS-4unR1 primer combination, while all other taxa, with the exception of Ericales, are decreasing in read abundance (Appendices 4 and 5).

The number of fungi sequences varies from < 1% to 80% of all reads in mix 1 and from < 1% to 88% in mix 2 (Table 3). Abundances are associated with the average number of mismatches in fungi (Tables 2 and 3). The primer combinations UniPlantF + UniPlantR, ITS-3p62plF1 + UniPlantR, ITS-p3 + ITS-p4 and ITS-S2F + BEL-3 are all expected to have three or more mismatch-

es in fungi which resulted in less than 1% of reads to be classified as fungi (Tables 2 and 3). The primer combinations UniPlantF + ITS-4unR1, ITS-u3 + ITS-u4 and ITS3 + ITS4 have, on average, less than one mismatch in fungi and produce fungal read abundances of 41% to 87% (Tables 2 and 3). The Lycopodiopsida in mix 1 are represented by 0 to 345 reads per 1,000 reads (Table 3; Appendices 4 and 5).

### Mock community PCR optimisations

Addition of 5% DMSO to the PCR mix has a threefold effect: 1) In most cases, it reduces the number of reads necessary to detect the species with a GC content of  $\geq 62\%$  (Fig. 2). 2) It reduces the number of taxa that have at least one failed replicate from nine to three for mix 1 and from six to three in mix 2 (Fig. 2). 3) It enables *Lindera* (mix 1) and *Asimina* (mix 2) to be detected (Fig. 2). Taxa like

*Smilax* (mix 1) and *Quercus* (mix 2) that did not show a substantial improvement compared to 0% DMSO, had an overall very low number of reads (Appendices 4 and 5).

## Discussion

### Selecting and refining: one possible approach to plant metabarcoding studies

Although universal ITS primers have been proposed in the past (White et al. 1990; Blattner 1999; Cheng et al. 2016; Moorhouse-Gann et al. 2018), the increasing availability of publicly available ITS sequences allows for an improvement in universality that was not previously possible. This is demonstrated by comparing the number of sequences (Cycadopsida, Gnetopsida, Pinopsida, Liliopsida and Magnoliopsida) used for primer design by Cheng et al. (2016) and this study (55,700 and 187,522 5.8S nrDNA sequences, respectively). Moorhouse-Gann et al. (2018) used a geographically restricted (Mauritius and United Kingdom) database of less than 10,000 5.8S nrDNA sequences for primer design.

A plant-specific primer with a low number of mismatches, weak secondary structures and a uniform amplification of a complex DNA mixture has the highest chance to deliver representative results when used in combination with an unknown eDNA sample (Tables 2 and 3). The detailed mismatch lists in this study furthermore allow primers to be customised to fit the exact needs of a given study (Suppl. material 1: Suppl. file 2 and Suppl. material 1: Suppl. file 3). Especially studies including Orchidaceae may find that their unique sequence characteristics warrant extra attention (Suppl. material 1: Suppl. file 3). With the help of the family-level alignments (Suppl. material 1: Suppl. file 3), an orchid-specific primer can be synthesised and added to the degenerate primer mix in a relative quantity of  $1/n$  ( $n$  = total number of primer variations of the respective degenerate primer).

### In silico primer evaluation

Finding the balance between the elimination of primer mismatches and the number of total primer variants was one of the main design goals of this study. The five primers we generated (ITS-2pLR1, ITS-3p\* and ITS-4unR1) achieved less mismatches than most of the previously published ITS primers (Table 2). This results are a significant improvement, as it is rather difficult to predict whether or not a mismatch is critically affecting subsequent amplification or may be of minor importance. The impact of primer mismatches on the overall assay success are determined by parameters that include, but are not limited to, salt concentration, annealing temperature, mismatch position, mispaired nucleotide and template concentrations (Ayyadevara et al. 2000; Waterfall et al. 2002; Sipos et al. 2007; Wu et al. 2009; Lefever et al. 2013; Rejali et al. 2018). Studies by Lefever et al. (2013)

show that the impact of (non-3'-terminal) primer mismatches on primer melting temperature is hard to predict, ranging from 0–8 °C for one mismatch and 2–20 °C for two mismatches. Despite the fact that PCR performance has been reported to decrease fundamentally in severity if the mismatch is located more than 8 bp (Rejali et al. 2018) or more than 12 bp (Wu et al. 2009) away from the 3' terminal position, multiple studies suggest that mismatches towards the 5' end of primers may not be completely neutral (Sipos et al. 2007; Boyle et al. 2009; Lefever et al. 2013). For these reasons, up-to-date primer development with a minimal number of mismatches is a prerequisite for any successful DNA-based environmental assay, as it increases the number of true positive detections.

Although the study by Cheng et al. (2016) shows a drastic reduction of mismatches of more than 80% in Angiosperms in the ITS-u3 primer versus the ITS3 primer, our results indicate only a reduction of approximately 15% (Table 2). This can be attributed to the way mismatches have been counted. This study counts mismatches and summarises them on a family level, while Cheng et al. (2016) counts mismatches on a sequence-based level. This leads to a scenario where 15% of the included Angiosperm plant families represent 50% of the informative nature of the analysis. We believe that our method of analysis is better suited for a wide breadth of applications, as families with little researched taxa are given the same weight as plant families with a better sequence coverage.

An additional advantage of our method is that it reveals mismatch patterns more easily. If a mismatch occurs consistently in a large number of families and if some of these families are only represented by few sequences, the underlying pattern becomes very hard to catch, if the analysis is based on sequence level only. This can be seen in the ITS-p4 and ITS-u2 primer (Suppl. material 1: Suppl. file 3). The majority of mismatches in these primers could have been eliminated by introducing a single ambiguous nucleotide (ITS-p4: 10G > K and ITS-u2: 10G > R). Making three modifications in the ITS-p3 primer (i.e. 8C>Y, 10G > R, 14C > Y) eliminates most of its mismatches. However, unfortunately, this would introduce a hairpin structure with a melting temperature above the primer melting temperature (Suppl. material 1: Suppl. file 5). In addition to the positioning of the ITS-p3 primer, which adds 140 bp of minimally informative nucleotides to the amplicon length, this limits the potential of the ITS-p3 primer.

The threshold we report of a 30% frequency of mismatch in the corresponding plant family prevents outdated taxonomic assignments (i.e. a genus is placed in the wrong family) and misidentified sequences from introducing noise, which would result in primer design with an unnecessarily high number of ambiguous nucleotides. However, there are rare cases of mismatches of the same nucleotide and at the same primer position that occur at very low frequency within a noticeable number of plant families (Suppl. material 1: Suppl. file 2). One example shown in our data is the UniPlantF and the (partially over-

lapping) ITS-3p62plF1 primer (Suppl. material 1: Suppl. file 2). For this reason, although Moorhouse-Gann et al. (2018) introduced an ambiguous nucleotide at primer position 12G > R, we did not replace 18G > R in the ITS-3p62plF1 primer due to the low prevalence of the mismatch as well as this modification resulting in the introduction of a hairpin structure. Due to the high melting temperature of the hairpin structure, which is close to the primer annealing temperature, there is a high risk that the gained universality will not translate into increased PCR success in the targeted plant families. Furthermore, this modification would introduce additional seven new primer variants that do not match any Spermatophyta template (Suppl. material 1: Suppl. file 5). There is also a geographical aspect, as the affected plant families generally have their centre of diversity in tropical regions (Köppen-Geiger climate classification: A). This illustrates that even primers optimised for universality require careful evaluation to find the best possible match between primer characteristics and the target of the respective study.

Strictly eliminating all mismatches by replacing each mismatch by an ambiguous nucleotide increases the count of total primer variations exponentially. Every introduced ambiguous nucleotide comes with a risk of negatively impacting the primer performance (e.g. primer dimers). An alternative for primers with ambiguous nucleotides that have already been optimised as far as possible, considering a reasonable number of total primer variations and still have mismatches in a few plant families, would be to add only those primer variations that lead to the correction of the respective mismatch. Our results allow researchers to supplement existing primers with fixed (without ambiguous nucleotides) primers, targeting specific mismatches of plant families to tailor the primer mixture to their individual needs (Suppl. material 1: Suppl. file 3).

A modification of the ITS-4unR1 primer (11T>D) allows flexibility to either be truly universal or exclude certain taxa, at least partially wind-pollinated plant families. This could be useful in insect pollination studies to reduce the number of reads generated by NGS from some *Pinus* species (e.g. *Pinus sylvestris*).

Like previous studies (Cheng et al. 2016; Moorhouse-Gann et al. 2018), this analysis is limited by the quantity of publicly available of ITS sequences and a future re-evaluation of the primers introduced here will be necessary to verify their universality. Filling the taxonomic gaps in public sequence repositories is of high importance, as reliable biodiversity monitoring via metabarcoding cannot be achieved without a complete regional reference database. Currently, there is no plant barcode marker available which covers more than 25% of all plant species, known or unknown (Corlett 2016; Kolter and Gemeinholzer 2020).

### Mock community primer test

Due to PCR stochasticity (Kebschull and Zador 2015) and nrDNA copy number variation between 150 and

> 100,000 per genome (Prokopowich et al. 2003; Wang et al. 2019) sequencing results of the mock community should be treated qualitatively and not quantitatively. Therefore, we scored the number of missed taxa instead of their relative abundance (read counts) as a measure of primer fitness. The minimally-required read depth provides an indication of how evenly amplification occurred (Table 3). We purposefully did not correct for the differences in nrDNA copy numbers to create mock communities in which some taxa are under-represented by at least one order of magnitude.

Having these aforementioned restrictions in mind, selection of the best suited primers is, in general, the most important tool in minimising additional biases during PCR and library preparation (Schirmer et al. 2015). Although previous studies have focused on optimising existing ITS primers or generating new ones (Morgan et al. 2009; Cheng et al. 2016; Moorhouse-Gann et al. 2018), none of them evaluated primer performance using a mock community. We effectively address this knowledge gap by the first primer efficacy assessment, based on two plant mock communities. The importance of this approach is underlined by the fact that some primers selectively disfavour a certain set of taxa, even without any mismatch being present. Other possible explanations for primer specific taxon bias are secondary structures or stretches with high or low GC content close or at the primer binding site and different binding characteristics of the variants of a primer with ambiguous nucleotides. This may be the case for *Gagea graeca* (Appendices 4 and 5), as it is nearly absent from sequencing runs using the UniPlantR primer, which has no mismatch in *Gagea graeca*. In addition to its bias, the mismatches of the UniPlantR primer in other Liliopsida families warrant caution when using this primer in combination with unknown eDNA samples (Suppl. material 1: Suppl. file 3).

One possible explanation for the mixed performance of the 58SPL primer could be that the relatively high primer hairpin melting temperature makes a large proportion of the 58SPL primer unavailable to anneal to its intended template sequence, disfavours amplicons with a rare prevalence (Table 2). If this is the case, increasing the primer concentration or raising the annealing temperature might alleviate this issue.

The ITS3 + ITS4 primer combination, included in this study, was originally designed to amplify fungi (White et al. 1990). Due to its confirmed high preference towards fungi, as well as the high missed taxa rate, we discourage the use of this primer pair for plant metabarcoding. In spite of this, the ITS3 + ITS4 primer combination still holds its place in recent literature (Gresty et al. 2018; Kamo et al. 2018; Besse 2021; Câmara et al. 2021).

The ITS-u3 + ITS-u4, ITS-p3 + ITS-p4 and ITS-S2F + BEL-3 primer combinations have some performance metrics in their favour (Table 3). Despite this, the relatively high number of plant families with mismatches, as shown by the *in silico* analysis (Table 2), warrants careful evaluation before using them with eDNA. These results

emphasise that, despite the usefulness of mock communities, *in silico* evaluation can provide valuable additional information on primer universality. While the universality of aforementioned primers could be improved by adding additional ambiguities, the overlap with already existing or already optimised primers, introduced in this paper, indicates that most of them can be replaced with updated versions (Fig. 1).

To our knowledge, this is the first mock community-based primer comparison in the context of metabarcoding in the plant kingdom. Although our mock communities only reflect a fraction of the genetic diversity within plants, we have demonstrated that there are differences between different ITS primer combinations and that these differences are not necessarily based on primer mismatches. The differences not originating from primer mismatches cannot be detected by *in silico* analyses only, further illustrating the need for mock community studies to verify the results of metabarcoding programmes. In contrast, the universality of the primers can be better assessed by the more comprehensive evaluations made during the *in silico* analysis. Should the need arise, in essence, we recommend an integrative approach to evaluate primers by combining *in silico* and mock community analyses. The composition of the mock community should ideally be connected to the respective study area. Considering that we thoroughly screened the whole 5.8S nrDNA region for potential primer sequences, we advise using one of the primers presented in this paper as a starting point for further refinement, if needed.

### Mock community PCR optimisations

Varadharajan and Parani (2021) mentions 65% GC content as the threshold for which additives are impacting the PCR success dramatically; similarly, our results indicate 62% GC as the threshold where DMSO starts to improve the sequencing result of a diverse mock community (Fig. 2). In accordance with Varadharajan and Parani (2021), we also show that concentrations below 5% DMSO did not suffice to maximise amplification success (Suppl. material 1: Suppl. file 1). Of the ITS2 plant sequences available in public repositories in 2018, 28.5% had a GC content of 65% or higher (Kolter and Gemeinholzer 2020). The lack of PCR optimisation for high GC levels may invalidate the main findings of previous studies, as plants with high GC levels are eliminated from the amplicon pool during PCR.

## Conclusion

In metabarcoding, regardless of the marker used, we point out that it is strongly recommended to integrate mock communities into the workflow to provide additional quality control (Thalinger et al. 2020). Based on the results of our *in silico* and mock community analyses, we recommend ITS-3p62pIF1 in combination with

ITS-4unR1 (possibly modified) for amplification of the ITS2 region. The UniPlantF + ITS-p4 (modified) needs additional validation, but shows promise to also improve future ITS2 metabarcoding studies. If sequence length is decisive for primer selection, the SPL58 primer, in combination with the UniPlantR primer, offers the shortest possible ITS2 amplicon. However, secondary structures and mismatches could negatively affect PCR efficiency and universality.

The past has shown that ITS-based studies have struggled with amplification success (Braukmann et al. 2017; Gill et al. 2019; Jones et al. 2021). However, this work eliminates the most pressing issues, namely the lack of PCR optimisations and the lack of a comprehensive primer evaluation. In contrast to other plant markers, the ITS region now combines high informative and the potential for high amplification success.

## Acknowledgements

We thank Volker Wissemann for access to the lab and Martin de Jong for helping with the acquisition of plant material.

## Funding statement

This work has been funded by the DFG SPP1991 priority programme “Taxon-Omics: New approaches for discovering and naming biodiversity”.

## Ethics statement

The authors declare no conflict of interest.

## References

- Alanagreh L, Pegg C, Harikumar A, Buchheim M (2017) Assessing intragenomic variation of the internal transcribed spacer two: Adapting the Illumina metagenomics protocol. *PLoS ONE* 12: e0181491. <https://doi.org/10.1371/journal.pone.0181491>
- Allawi HT, SantaLucia J (1997) Thermodynamics and NMR of internal G.T mismatches in DNA. *Biochemistry* 36: 10581–10594. <https://doi.org/10.1021/bi962590c>
- Ankenbrand MJ, Keller A, Wolf M, Schultz J, Förster F (2015) ITS2 Database V: Twice as Much. *Molecular Biology and Evolution* 32: 3030–3032. <https://doi.org/10.1093/molbev/msv174>
- Antonelli A, Fry C, Smith RJ, Simmonds M, Kersey PJ, Pritchard HW, Abbo MS, Acedo C, Adams J, Ainsworth AM, Allkin B, Annecke W, Bachman SP, Bacon K, Bárríos S, Barstow C, Battison A, Bell E, Bensusan K, Bidartondo MI, Blackhall-Miles RJ, Borrell JS, Brearley FQ, Breman E, Brewer R, Brodie J, Cámara-Leret R, Campostrini Forzza R, Cannon P, Carine M, Carretero J, Cavagnaro TR, Cazar M-E, Chapman T, Cheek M, Clubbe C, Cockel C, Collemare

- J, Cooper A, Copeland AI, Corcoran M, Couch C, Cowell C, Crous P, Da Silva M, Dalle G, Das D, David JC, Davies L, Davies N, Canha MN de, Lirio EJ de, Demissew S, Diazgranados M, Dickie J, Dines T, Douglas B, Dröge G, Dulloo ME, Fang R, Farlow A, Farrar K, Fay MF, Felix J, Forest F, Forrest LL, Fulcher T, Gafforov Y, Gardiner LM, Gâteblé G, Gaya E, Geslin B, Gonçalves SC, Gore C, Govaerts R, Gowda B, Grace OM, Grall A, Haelewaters D, Halley JM, Hamilton MA, Hazra A, Heller T, Hollingsworth PM, Holstein N, Howes M-J, Hughes M, Hunter D, Hutchinson N, Hyde K, Iganci J, Jones M, Kelly LJ, Kirk P, Koch H, Grisai-Greilhuber I, Lall N, Langat MK, Leaman DJ, Leão TC, Lee MA, Leitch IJ, Leon C, Lettice E, Lewis GP, Li L, Lindon H, Liu JS, Liu U, Llewellyn T, Looney B, Lovett JC, Luczaj L, Lulekal E, Maggassouba S, Malécot V, Martin C, Masera OR, Mattana E, Maxted N, Mba C, McGinn KJ, Metherringham C, Miles S, Miller J, Milliken W, Moat J, Moore P, Morim MP, Mueller GM, Muminjanov H, Negrão R, Nic Lughadha E, Nicholson N, Niskanen T, Nono Womdim R, Noorani A, Obreza M, O'Donnell K, O'Hanlon R, Onana J-M, Ondo I, Padulosi S, Paton A, Pearce T, Pérez Escobar OA, Pieroni A, Pironon S, Prescott T, Qi YD, Qin H, Quave CL, Rajaovelona L, Razanajatovo H, Reich PB, Rianawati E, Rich T, Richards SL, Rivers MC, Ross A, Rumsey F, Ryan M, Ryan P, Sagala S, Sanchez MD, Sharrock S, Shrestha KK, Sim J, Sirakaya A, Sjöman H, Smidt EC, Smith D, Smith P, Smith SR, Sofo A, Spence N, Stanworth A, Stara K, Stevenson PC, Stroh P, Suz LM, Tambam BB, Tatsis EC, Taylor I, Thiers B, Thormann I, Vaglica V, Vásquez-Londoño C, Victor J, Viruel J, Walker BE, Walker K, Walsh A, Way M, Wilbraham J, Wilkin P, Wilkinson T, Williams C, Winterton D, Wong KM, Woodfield-Pascoe N, Woodman J, Wyatt L, Wynberg R, Zhang BG (2020) State of the World's Plants and Fungi 2020. Royal Botanic Gardens, Kew.
- Ayyadevara S, Thaden JJ, Shmookler Reis RJ (2000) Discrimination of primer 3'-nucleotide mismatch by taq DNA polymerase during polymerase chain reaction. *Analytical Biochemistry* 284: 11–18. <https://doi.org/10.1006/abio.2000.4635>
- Besse P (2021) *Molecular Plant Taxonomy*. Springer US, New York, [xiii +] 394 pp. <https://doi.org/10.1007/978-1-0716-0997-2>
- Bjørnsgaard AA, Davey ML, Kauserud H (2017) ITS all right mama: investigating the formation of chimeric sequences in the ITS2 region by DNA metabarcoding analyses of fungal mock communities of different complexities. *Molecular Ecology Resources* 17: 730–741. <https://doi.org/10.1111/1755-0998.12622>
- Blattner FR (1999) Direct amplification of the entire ITS region from poorly preserved plant material using recombinant PCR. *BioTechniques* 27: 1180–1186. <https://doi.org/10.2144/99276st04>
- Boyle B, Dallaire N, MacKay J (2009) Evaluation of the impact of single nucleotide polymorphisms and primer mismatches on quantitative PCR. *BMC Biotechnology* 9: e75. <https://doi.org/10.1186/1472-6750-9-75>
- Braukmann TWA, Ivanova NV, Prosser SWJ, Elbrecht V, Steinke D, Ratnasingham S, Waard JR de, Sones JE, Zakharov EV, Hebert PDN (2019) Metabarcoding a diverse arthropod mock community. *Molecular Ecology Resources* 19: 711–727. <https://doi.org/10.1111/1755-0998.13008>
- Braukmann TWA, Kuzmina ML, Sills J, Zakharov EV, Hebert PDN (2017) Testing the efficacy of DNA Barcodes for identifying the vascular plants of Canada. *PLoS ONE* 12: e0169515. <https://doi.org/10.1371/journal.pone.0169515>
- Câmara PEAS, Carvalho-Silva M, Pinto OHB, Amorim ET, Henriques DK, Da Silva TH, Pellizzari F, Convey P, Rosa LH (2021) Diversity and Ecology of Chlorophyta (Viridiplantae) Assemblages in protected and non-protected sites in Deception Island (Antarctica, South Shetland Islands) Assessed Using an NGS Approach. *Microbial Ecology* 81: 323–334. <https://doi.org/10.1007/s00248-020-01584-9>
- Chamberlain SA, Boettiger C (2017) R Python, and Ruby clients for GBIF species occurrence data. *PeerJ Preprints*. <https://doi.org/10.7287/peerj.preprints.3304v1>
- Chen S, Yao H, Han J, Liu C, Song J, Shi L, Zhu Y, Ma X, Gao T, Pang X, Luo K, Li Y, Li X, Jia X, Lin Y, Leon C (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS ONE* 5: e8613. <https://doi.org/10.1371/journal.pone.0008613>
- Cheng T, Xu C, Lei L, Li C, Zhang Y, Zhou S (2016) Barcoding the kingdom Plantae: new PCR primers for ITS regions of plants with improved universality and specificity. *Molecular Ecology Resources* 16: 138–149. <https://doi.org/10.1111/1755-0998.12438>
- Chiou S-J, Yen J-H, Fang C-L, Chen H-L, Lin T-Y (2007) Authentication of medicinal herbs using PCR-amplified ITS2 with specific primers. *Planta Medica* 73: 1421–1426. <https://doi.org/10.1055/s-2007-990227>
- Corlett RT (2016) Plant diversity in a changing world: Status, trends, and conservation needs. *Plant Diversity* 38: 10–16. <https://doi.org/10.1016/j.pld.2016.01.001>
- Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, Vere N de, Pfrender ME, Bernatchez L (2017) Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology* 26: 5872–5895. <https://doi.org/10.1111/mec.14350>
- Edgar RC (2016) SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv*. <https://doi.org/10.1101/074161>
- Elbrecht V, Braukmann TWA, Ivanova NV, Prosser SWJ, Hajibabaei M, Wright M, Zakharov EV, Hebert PDN, Steinke D (2019) Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ* 7: e7745. <https://doi.org/10.7717/peerj.7745>
- Elbrecht V, Leese F (2017) Validation and development of COI metabarcoding primers for freshwater macroinvertebrate bioassessment. *Frontiers in Environmental Science* 5: e11. <https://doi.org/10.3389/fenvs.2017.00011>
- Fazekas AJ, Kuzmina ML, Newmaster SG, Hollingsworth PM (2012) DNA barcoding methods for land plants. *Methods in Molecular Biology* (Clifton, N.J.) 858: 223–252. [https://doi.org/10.1007/978-1-61779-591-6\\_11](https://doi.org/10.1007/978-1-61779-591-6_11)
- Feliner GN, Rosselló JA (2007) Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. *Molecular Phylogenetics And Evolution* 44: 911–919. <https://doi.org/10.1016/j.ympev.2007.01.013>
- Gill BA, Musili PM, Kurukura S, Hassan AA, Goheen JR, Kress WJ, Kuzmina M, Pringle RM, Kartzinel TR (2019) Plant DNA-barcode library and community phylogeny for a semi-arid East African savanna. *Molecular Ecology Resources* 19: 838–846. <https://doi.org/10.1111/1755-0998.13001>
- Glenn TC, Pierson TW, Bayona-Vásquez NJ, Kieran TJ, Hoffberg SL, Thomas IV JC, Lefever DE, Finger JW, Gao B, Bian X, Louha S, Kolli RT, Bentley KE, Rushmore J, Wong K, Shaw TI, Rothrock MJ, McKee AM, Guo TL, Mauricio R, Molina M, Cummings BS, Lash LH, Lu K, Gilbert GS, Hubbell SP, Faircloth BC (2019) Adapterama

- II: universal amplicon sequencing on Illumina platforms (TaggIMatrix). *PeerJ* 7: e7786. <https://doi.org/10.7717/peerj.7786>
- Gresty CEA, Clare E, Devey DS, Cowan RS, Csiba L, Malakasi P, Lewis OT, Willis KJ (2018) Flower preferences and pollen transport networks for cavity-nesting solitary bees: Implications for the design of agri-environment schemes. *Ecology and Evolution* 8: 7574–7587. <https://doi.org/10.1002/ece3.4234>
- Grooten M, Almond RE, Peterson T [Eds] (2020) Living Planet Report 2020. Bending the curve of biodiversity loss, 1<sup>st</sup> edn. World-Wide Fund for Nature, Gland, Switzerland.
- Gu W, Song J, Cao Y, Sun Q, Yao H, Wu Q, Chao J, Zhou J, Xue W, Duan J (2013) Application of the ITS2 region for barcoding medicinal plants of Selaginellaceae in Pteridophyta. *PLoS ONE* 8: e67818. <https://doi.org/10.1371/journal.pone.0067818>
- Hajibabaei M, Porter TM, Wright M, Rudar J (2019) COI metabarcoding primer choice affects richness and recovery of indicator taxa in freshwater systems. *PLoS ONE* 14: e0220953. <https://doi.org/10.1371/journal.pone.0220953>
- IPBES (2019) Summary for policymakers of the global assessment report on biodiversity and ecosystem services. Zenodo.
- Jones L, Twyford AD, Ford CR, Rich TCG, Davies H, Forrest LL, Hart ML, McHaffie H, Brown MR, Hollingsworth PM, Vere N de (2021) Barcode UK: A complete DNA barcoding resource for the flowering plants and conifers of the United Kingdom. *Molecular Ecology Resources* 21(6): 2050–2062. <https://doi.org/10.1111/1755-0998.13388>
- Kamo T, Kusumoto Y, Tokuoka Y, Okubo S, Hayakawa H, Yoshiyama M, Kimura K, Konuma A (2018) A DNA barcoding method for identifying and quantifying the composition of pollen species collected by European honeybees, *Apis mellifera* (Hymenoptera: Apidae). *Applied Entomology and Zoology* 53: 353–361. <https://doi.org/10.1007/s13355-018-0565-9>
- Kay KM, Whittall JB, Hodges SA (2006) A survey of nuclear ribosomal internal transcribed spacer substitution rates across angiosperms: an approximate molecular clock with life history effects. *BMC Evolutionary Biology* 6: e36. <https://doi.org/10.1186/1471-2148-6-36>
- Kebschull JM, Zador AM (2015) Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Research* 43: e143. <https://doi.org/10.1093/nar/gkv717>
- Kelly RP, Shelton AO, Gallego R (2019) Understanding PCR processes to draw meaningful conclusions from environmental DNA studies. *Scientific Reports* 9: e12133. <https://doi.org/10.1038/s41598-019-48546-x>
- Kolter A, Gemeinholzer B (2020) Plant DNA barcoding necessitates marker-specific efforts to establish more comprehensive reference databases. *Genome*. <https://doi.org/10.1139/gen-2019-0198>
- Krehenwinkel H, Wolf M, Lim JY, Rominger AJ, Simison WB, Gillespie RG (2017) Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Scientific Reports* 7: e17668. <https://doi.org/10.1038/s41598-017-17333-x>
- Lefever S, Pattyn F, Hellemans J, Vandesompele J (2013) Single-nucleotide polymorphisms and other mismatches reduce performance of quantitative PCR assays. *Clinical Chemistry* 59: 1470–1480. <https://doi.org/10.1373/clinchem.2013.203653>
- Li S, Deng Y, Wang Z, Zhang Z, Kong X, Zhou W, Yi Y, Qu Y (2020) Exploring the accuracy of amplicon-based internal transcribed spacer markers for a fungal community. *Molecular Ecology Resources* 20: 170–184. <https://doi.org/10.1111/1755-0998.13097>
- Liu Z-W, Gao Y-Z, Zhou J (2019) Molecular Authentication of the medicinal species of *Ligusticum* (*Ligustici Rhizoma et Radix*, “Gaoben”) by integrating non-coding Internal Transcribed Spacer 2 (ITS2) and its secondary structure. *Frontiers in Plant Science* 10: e429. <https://doi.org/10.3389/fpls.2019.00429>
- Marsh DM, Trenham PC (2008) Current trends in plant and animal population monitoring. *Conservation biology: the Journal of the Society for Conservation Biology* 22: 647–655. <https://doi.org/10.1111/j.1523-1739.2008.00927.x>
- Moorhouse-Gann RJ, Dunn JC, Vere N de, Goder M, Cole N, Hipperperson H, Symondson WOC (2018) New universal ITS2 primers for high-resolution herbivory analyses using DNA metabarcoding in both tropical and temperate zones. *Scientific Reports* 8: e8542. <https://doi.org/10.1038/s41598-018-26648-2>
- Morgan M, Anders S, Lawrence M, Aboyoun P, Pagès H, Gentleman R (2009) ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics (Oxford, England)* 25: 2607–2608. <https://doi.org/10.1093/bioinformatics/btp450>
- Owczarzy R, Tataurov AV, Wu Y, Manthey JA, McQuisten KA, Almabrazi HG, Pedersen KF, Lin Y, Garretson J, McEntaggart NO, Sailor CA, Dawson RB, Peek AS (2008) IDT SciTools: a suite for analysis and design of nucleic acid oligomers. *Nucleic Acids Research* 36: W163–169. <https://doi.org/10.1093/nar/gkn198>
- Pagès HA, Gentleman P, DebRoy R (2020) S. R package version 2.56.0. Biostrings: Efficient manipulation of biological strings.
- Piñol J, Senar MA, Symondson WOC (2019) The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. *Molecular Ecology* 28: 407–419. <https://doi.org/10.1111/mec.14776>
- Piper AM, Batovska J, Cogan NOI, Weiss J, Cunningham JP, Rodoni BC, Blacket MJ (2019) Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance. *GigaScience* 8(8): 1–22. <https://doi.org/10.1093/gigascience/giz092>
- Porter TM, Hajibabaei M (2018) Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Molecular Ecology* 27: 313–338. <https://doi.org/10.1111/mec.14478>
- Prokopowich CD, Gregory TR, Crease TJ (2003) The correlation between rDNA copy number and genome size in eukaryotes. *Genome* 46: 48–50. <https://doi.org/10.1139/g02-103>
- Prosekov A, Kuznetsov A, Rada A, Ivanova S (2020) Methods for monitoring large terrestrial animals in the wild. *Forests* 11(8): e808. <https://doi.org/10.3390/f11080808>
- R Core Team (2020) R: A Language and Environment for Statistical Computing, Vienna.
- Rejali NA, Moric E, Wittwer CT (2018) The effect of single mismatches on primer extension. *Clinical Chemistry* 64: 801–809. <https://doi.org/10.1373/clinchem.2017.282285>
- Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4: e2584. <https://doi.org/10.7717/peerj.2584>
- Ruppert KM, Kline RJ, Rahman MS (2019) Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation* 17: e00547. <https://doi.org/10.1016/j.gecco.2019.e00547>
- Schirmer M, Ijaz UZ, D’Amore R, Hall N, Sloan WT, Quince C (2015) Insight into biases and sequencing errors for amplicon sequencing

with the Illumina MiSeq platform. *Nucleic Acids Research* 43: e37. <https://doi.org/10.1093/nar/gku1341>

Sellers GS, Di Muri C, Gómez A, Hänfling B (2018) Mu-DNA: a modular universal DNA extraction method adaptable for a wide range of sample types. *Metabarcoding and Metagenomics* 2: e24556. <https://doi.org/10.3897/mbmg.2.24556>

Sipos R, Székely AJ, Palatinszky M, Révész S, Márialigeti K, Nikolausz M (2007) Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targeting bacterial community analysis. *FEMS Microbiology Ecology* 60: 341–350. <https://doi.org/10.1111/j.1574-6941.2007.00283.x>

Smith KF, Kohli GS, Murray SA, Rhodes LL (2017) Assessment of the metabarcoding approach for community analysis of benthic-epiphytic dinoflagellates using mock communities. *New Zealand Journal of Marine and Freshwater Research* 51: 555–576. <https://doi.org/10.1080/00288330.2017.1298632>

Song J, Shi L, Li D, Sun Y, Niu Y, Chen Z, Luo H, Pang X, Sun Z, Liu C, Lv A, Deng Y, Larson-Rabin Z, Wilkinson M, Chen S (2012) Extensive pyrosequencing reveals frequent intra-genomic variations of internal transcribed spacer regions of nuclear ribosomal DNA. *PLoS ONE* 7: e43971. <https://doi.org/10.1371/journal.pone.0043971>

Thalinger B, Deiner K, Harper LR, Rees HC, Blackman RC, Sint D, Traugott M, Goldberg CS, Bruce K (2020) A validation scale to determine the readiness of environmental DNA assays for routine species monitoring. *bioRxiv*. <https://doi.org/10.1101/2020.04.27.063990>

Tremblay ÉD, Duceppe M-O, Thurston GB, Gagnon M-C, Côté M-J, Bilodeau GJ (2019) High-resolution biomonitoring of plant pathogens and plant species using metabarcoding of pollen pellet contents collected from a honey bee hive. *Environmental DNA* 1: 155–175. <https://doi.org/10.1002/edn3.17>

Varadharajan B, Parani M (2021) DMSO and betaine significantly enhance the PCR amplification of ITS2 DNA barcodes from plants. *Genome* 64: 165–171. <https://doi.org/10.1139/gen-2019-0221>

Wang W, Wan T, Becher H, Kuderova A, Leitch IJ, Garcia S, Leitch AR, Kovařík A (2019) Remarkable variation of ribosomal DNA organization and copy number in gnetophytes, a distinct lineage of gymnosperms. *Annals of Botany* 123: 767–781. <https://doi.org/10.1093/aob/mcy172>

Waterfall CM, Eiseenthal R, Cobb BD (2002) Kinetic characterisation of primer mismatches in allele-specific PCR: a quantitative assessment. *Biochemical and Biophysical Research Communications* 299: 715–722. [https://doi.org/10.1016/S0006-291X\(02\)02750-X](https://doi.org/10.1016/S0006-291X(02)02750-X)

White TJ, Bruns T, Lee S, Taylor J (1990) 38 – Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis MA, Gelfand DH, Sninsky JJ, White TJ (Eds) *PCR Protocols*: Elsevier, 315–322. <https://doi.org/10.1016/B978-0-12-372180-8.50042-1>

Wickham H (2016) ggplot2. *Elegant Graphics for Data Analysis*, 2<sup>nd</sup> edn. <https://doi.org/10.1007/978-3-319-24277-4>

Wu J-H, Hong P-Y, Liu W-T (2009) Quantitative effects of position and type of single mismatch on single base primer extension. *Journal of Microbiological Methods* 77: 267–275. <https://doi.org/10.1016/j.mimet.2009.03.001>

Zhang GK, Chain FJJ, Abbott CL, Cristescu ME (2018) Metabarcoding using multiplexed markers increases species detection in complex zooplankton communities. *Evolutionary Applications* 11: 1901–1914. <https://doi.org/10.1111/eva.12694>

## Appendix 1

**Table A1.** Taxonomic breakdown of sequences used for the respective in silico primer evaluation.

nrDNA region	sequences (min/max)	species (min/max)	families (min/max)	order (min/max)
SSU	22,574 / 63,024	15,180 / 35,113	95 / 149	38 / 44
5.8S*	187,522	85,362	210	53
LSU	25,845 / 90,319	14,663 / 43,795	109 / 176	42 / 49

\* partial 5.8S nrDNA sequences filtered prior to analysis.

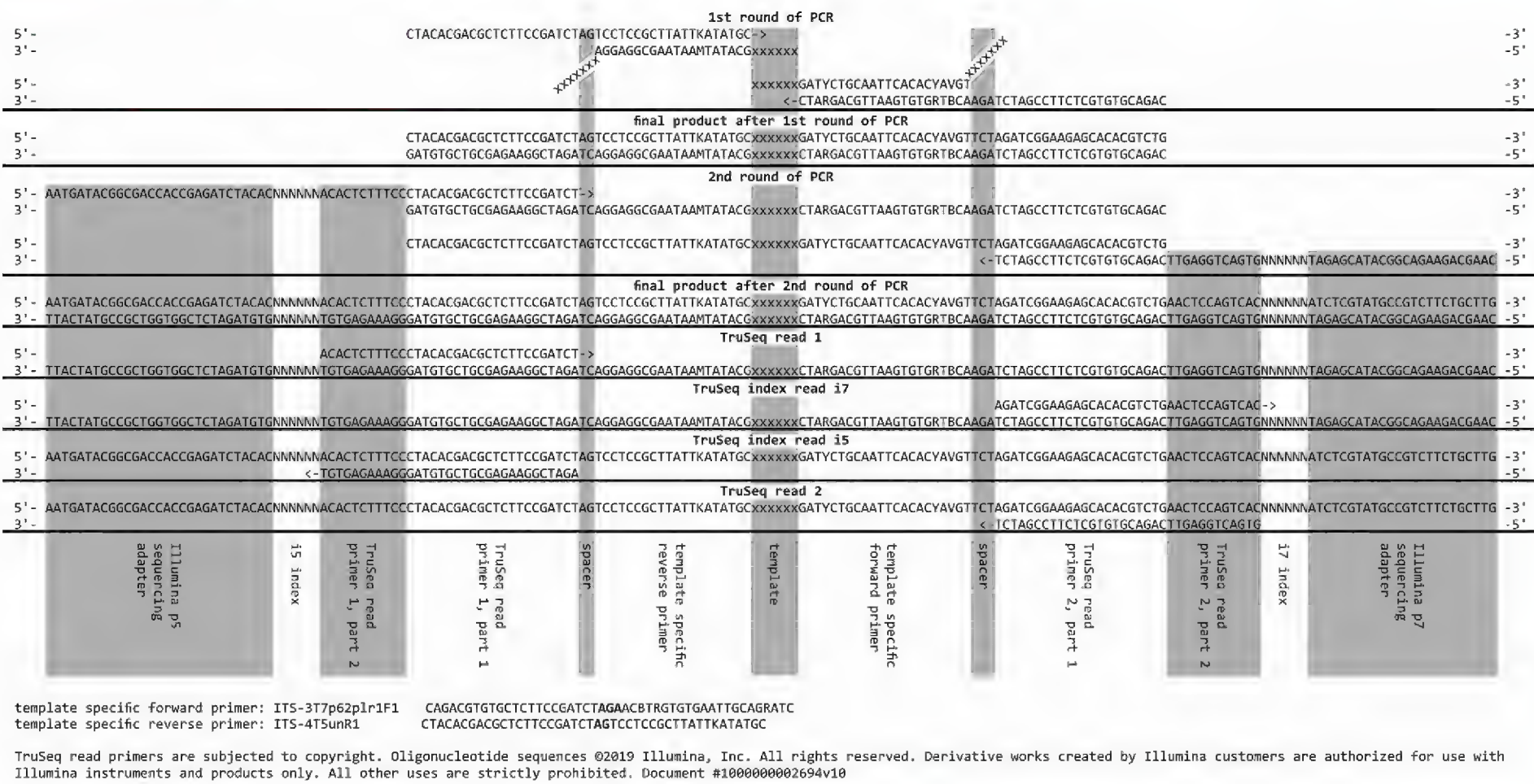
## Appendix 2

**Table A2.** PCR and sequencing primers used in MiSeq sequencing (5'→3').

location	orientation	unmodified primer name	modified primer name	primer tail	spacer	primer sequence
5.8S	forward	ITS3	ITS-3T73F1	CAGACGTGTGCTCTTCCGATCT		GCATCGATGAAGAACGCAGC
		ITS-S2F	ITS-S2F-T7	CAGACGTGTGCTCTTCCGATCT		ATGCGATACTTGGTGTGAAT
		ITS-u3	ITS-u3-T7	CAGACGTGTGCTCTTCCGATCT		CAWCGATGAAGAACGYAGC
		ITS-p3	ITS-p3-T7	CAGACGTGTGCTCTTCCGATCT		YGACTCTCGCAACGGATA
		UniPlantF	UniPlantF-T7	CAGACGTGTGCTCTTCCGATCT		TGTGAATTGCARRATYCMG
		ITS-3p62plF1	ITS-3T7p62plr1F1	CAGACGTGTGCTCTTCCGATCT	AGA	ACBTRGTGTGAATTGCAGRATC
			ITS-3T7p62plr2F1	CAGACGTGTGCTCTTCCGATCT	ATC	ACBTRGTGTGAATTGCAGRATC
			ITS-3T7p62plr3F1	CAGACGTGTGCTCTTCCGATCT	GCG	ACBTRGTGTGAATTGCAGRATC
				CAGACGTGTGCTCTTCCGATCT	AG	TTTGAACGCAAGTTGCGCC
		58SPL	ITS-2Cote-T7	CAGACGTGTGCTCTTCCGATCT		CCCCHYTGAYYTGRGGTCDC
LSU	reverse	UniPlantR	UniPlantR-T5	CTACACGACGCTCTTCCGATCT		TCCTCCGCTTATTKATATGC
		ITS-4unR1	ITS-4T5unR1	CTACACGACGCTCTTCCGATCT	AG	TCCTCCGCTTATTGATATGC
		ITS4	ITS-4T5unR2	CTACACGACGCTCTTCCGATCT	AG	RGTTCTTTCTCCGCTTA
		ITS-u4	ITS-u4-T5	CTACACGACGCTCTTCCGATCT		CCGCTTAKTGATATGCTTAA
		ITS-p4	ITS-p4-T5	CTACACGACGCTCTTCCGATCT		GACGCTTCTCCAGACTACAAT
		BEL-3	S3R-T5	CTACACGACGCTCTTCCGATCT		

Note: The primer tail is a part of the TruSeq read primers. Throughout the paper primers will be addressed by their unmodified names.

Appendix 3



**Figure A1.** Sequencing primer and two-step PCR layout. Note: The template specific primers used in this example are for demonstration purposes only and vary in each unique PCR setup. Additional template (x) strands bending away from the primer sequence in the 1st PCR round demonstrate their non-complementarity.

Appendix 4

Table A3. Mix 1 read abundances and GC values.

classification			GC content	read abundances (per 1.000 reads, median of replicates) per primer combination and presence / absence per replicate									
class	order	family	genus species	[%]	ITS-3p62plF1 + ITS-4unR1	UniPlantF + UniPlantR	UniPlantF + ITS-4unR1	ITS-3p62plF1 + UniPlantR	58SPL + ITS-4unR1	ITS-u3 + ITS-u4	ITS-p3 + ITS-p4	ITS-S2F + BEL-3	ITS3 + ITS4
Liliopsida	Asparagales	Asparagaceae	<i>Bellevaia trifoliata</i>	70	3.85	0.30	1.75	0.73	0.91	1.00	1.26	1.67	0.75
	Liliales	Liliaceae	<i>Gagea graeca</i>	65	182	0.57	71.1	0.64	43.3	61.5	169	25.3	36.9
	Liliales	Smilacaceae	<i>Smilax aspera</i>	71	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.00
	Poales	Cyperaceae	<i>Cyperus esculentus</i>	76	0.32	0.00	0.11	0.04	0.00	0.11	0.21	0.09	0.00
			<i>Schoenus nigricans</i>	70	2.47	0.16	1.55	0.17	0.50	0.65	1.53	0.24	0.28
	Poales	Poaceae	<i>Briza maxima</i>	64	0.39	0.08	0.20	0.08	0.06	0.30	0.38	0.40	0.08
	Apiales	Apiaceae	<i>Tordylium apulum</i>	56	22.5	7.85	9.32	9.09	5.18	6.06	10.8	6.05	2.83
	Brassicales	Brassicaceae	<i>Arabis verna</i>	52	163	429	95.5	435	166	23.1	115	91.7	17.7
	Dipsacales	Caprifoliaceae	<i>Valerianella discoidea</i>	72	19.5	2.89	7.84	3.56	11.6	5.86	12.4	24.9	2.16
	Ericales	Ericaceae	<i>Arbutus spec.#</i>	59	222	413	134	410	305	59.0	135	208	35.3
Magnoliopsida	Ericales	Ericaceae	<i>Erica arborea</i>	56	1.19	4.31	0.91	4.96	1.50	0.29	1.15	0.18	0.16
	Fabales	Fabaceae	<i>Anthyllis circinnata</i>	55	232	53.9	163	53.5	300	60.4	136	336	25.8
	Gentianales	Rubiaceae	<i>Sherardia arvensis</i>	62	2.93	4.73	1.21	5.94	0.36	1.28	2.20	0.00	0.57
			<i>Cinnamomum aromaticum</i>	76	0.06	0.00	0.00	0.05	0.00	0.06	0.00	0.09	0.00
	Laurales	Lauraceae	<i>Lindera obtusiloba</i>	77	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Malpighiales	Euphorbiaceae	<i>Mercurialis annua</i>	58	46.0	8.22	27.1	7.82	34.4	12.1	23.1	16.5	13.8
	Piperales	Aristolochiaceae	<i>Aristolochia guichardii</i>	76	0.85	0.41	0.50	0.36	0.38	0.18	0.69	0.19	0.22
	Rosales	Rosaceae	<i>Prunus dulcis</i>	65	9.60	11.2	3.92	12.6	3.45	3.01	8.30	5.35	1.43
	Sapindales	Anacardiaceae	<i>Pistacia lentiscus</i>	55	32.2	57.0	18.6	58.8	34.7	6.84	19.0	12.8	2.81
	Solanales	Solanaceae	<i>Solanum citrullifolium</i>	77	0.56	0.19	0.21	0.42	0.26	0.16	0.18	0.09	0.07
Pinopsida	Pinales	Pinaceae	<i>Pinus strobus</i>	59	0.22	0.54	0.10	0.63	0.08	0.29	0.16	0.00	0.00
					sum of missed taxa (read abundance ≤ 0.1 or missing in any replicates)								
					3	7	6	5	6	6	6	7	9
classification				GC content [%]	ITS-3p62plF1 + ITS-4unR1	UniPlantF + UniPlantR	UniPlantF + ITS-4unR1	ITS-3p62plF1 + UniPlantR	58SPL + ITS-4unR1	ITS-u3 + ITS-u4	ITS-p3 + ITS-p4	ITS-S2F + BEL-3	ITS3 + ITS4
Lycopodiopsida				67*	48.7	0.00	32.8	0.00	2.37	139	360	257	53.3
Fungi				35–72	6.79	1.72	443	0.04	103	619	0.00	0.00	800

Note: \* This GC content is the average of *Equisetum* (73% GC), *Huperzia* (72%GC) and *Selaginella* (57% GC). # Equal parts of *Arbutus unedo* and *Arbutus andrachne*. The ITS-S2F + BEL-3 primer combination yielded only 3 replicates. The ITS-p3 + ITS-p4 primer combination yielded less reads than the other combinations. The highest three read abundance values in each column, including fungi, are highlighted (bold). Absence of a taxon in one or more replicates is represented by a grid pattern where each white cell represents one replicate the respective taxa could not be detected in. *Philodendron angustisectum* could not be detected.

Appendix 5

Table A4. Mix 2 read abundance and GC values.

classification			GC content [%]	read abundances (per 1.000 reads, median) per primer combination and presence / absence per replicate											
class	order family	genus species		ITS-3p62pIF1 + ITS-4unR1	UniPlantF + UniPlantR	UniPlantF + ITS-4unR1	ITS-3p62pIF1 + UniPlantR	58SPL + ITS-4unR1	ITS-u3 + ITS-u4	ITS-p3 + ITS-p4	ITS-S2F + BEL-3	ITS3 + ITS4			
Liliopsida	Asparagales Asparagaceae	<i>Maianthemum bifolium</i>	74	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00			
	Liliales Liliaceae	<i>Gagea graeca</i>	64	<b>183</b>	0.62	<b>64.3</b>	0.92	<b>45.6</b>	<b>64.4</b>	<b>234</b>	34.7	<b>32.1</b>			
	Poales Cyperaceae	<i>Schoenus nigricans</i>	70	2.48	0.16	1.52	0.16	0.49	0.46	2.70	0.38	0.23			
	Apiales Apiaceae	<i>Tordylium apulum</i>	56	28.9	9.94	10.1	10.8	6.67	7.39	19.4	8.99	3.03			
	Asterales Asteraceae	<i>Geropogon hybridus</i>	56	86.0	<b>171</b>	40.2	<b>173</b>	32.6	19.2	98.2	<b>64.8</b>	8.99			
	Brassicales Brassicaceae	<i>Arabis verna</i>	52	<b>201</b>	<b>560</b>	<b>101</b>	<b>535</b>	<b>210</b>	<b>25.2</b>	<b>187</b>	<b>158</b>	<b>18.0</b>			
	Caryophyllales Caryophyllaceae	<i>Silene colorata</i>	60	39.3	67.1	12.8	83.5	4.56	11.7	55.6	15.0	6.96			
Magnoliopsida	Caryophyllales Polygonaceae	<i>Polygonum arenastrum</i>	77	4.75	16.5	2.10	23.8	3.47	0.96	4.27	2.19	0.51			
	Ericales Ericaceae	<i>Calluna vulgaris</i>	62	1.32	3.94	1.04	3.70	2.15	0.33	0.71	0.38	0.14			
		<i>Erica spec.</i>	<b>56</b>	1.39	5.39	0.73	5.33	1.38	0.28	0.89	0.38	0.23			
	Fabales Fabaceae	<i>Anthyllis circinnata</i>	55	<b>356</b>	<b>103</b>	<b>224</b>	<b>98.2</b>	<b>472</b>	<b>101</b>	<b>308</b>	<b>653</b>	<b>35.4</b>			
	Fagales Fagaceae	<i>Quercus ithaburensis</i>	68	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00			
	Laurales Lauraceae	<i>Laurus nobilis</i>	75	0.29	0.00	0.09	0.05	0.00	0.04	0.18	0.22	0.00			
	Magnoliales Annonaceae	<i>Asimina triloba</i>	79	0.15	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00			
Pinopsida	Malpighiales Euphorbiaceae	<i>Mercurialis annua</i>	58	51.8	11.0	26.3	8.83	43.6	14.2	36.6	30.8	13.8			
	Rosales Rosaceae	<i>Prunus dulcis</i>	65	11.0	14.9	4.00	16.7	4.45	3.09	12.4	9.70	1.32			
	Solanales Convolvulaceae	<i>Convolvulus siculus</i>	58	25.7	30.3	8.60	37.8	4.29	7.06	35.3	17.2	3.84			
	Pinales Pinaceae	<i>Pinus strobus</i>	59	0.35	0.65	0.21	0.74	0.05	0.27	0.34	0.15	0.00			
				sum of missed taxa (read abundance ≤ 0.1 or missing in any replicates)											
				3	5	5	4	5	4	5	3	6			
classification			GC content [%]	ITS-3p62pIF1 + ITS-4unR1	UniPlantF + UniPlantR	UniPlantF + ITS-4unR1	ITS-3p62pIF1 + UniPlantR	58SPL + ITS-4unR1	ITS-u3 + ITS-u4	ITS-p3 + ITS-p4	ITS-S2F + BEL-3	ITS3 + ITS4			
Fungi			35–72	10.4	2.4	508	0.04	167	743	0.00	0.00	876			

Note: The ITS-S2F + BEL-3 primer combination yielded only 3 replicates. The ITS-p3 + ITS-p4 primer combination yielded less reads than the other combinations. The highest three read abundance values in each column, excluding fungi, are highlighted (bold). Absence of a taxon in one or more replicates is represented by a grid pattern where each white cell represents one replicate the respective taxa could not be detected in. *Sassafras albidum* could not be detected.

Appendix 6

Table A5. Minimal required read depth for mix 1.

classification			GC [%]	required minimal read depth to achieve a detection with 95% probability (no singletons) [thousands]									
class	order	family		ITS-3p62p1F1 + ITS-4unR1	UniPlantF + UniPlantR	UniPlantF + ITS-4unR1	ITS-3p62p1F1 + UniPlantR	58SPL + ITS-4unR1	ITS-u3 + ITS-u4	ITS-p3 + ITS-p4	ITS-S2F + BEL-3	ITS3 + ITS4	
Liliopsida	Asparagales	Asparagaceae	70	1.3	NA		2.8	7	5	5.2	3.7	2.5	6
	Liliales	Liliaceae	65	0.1	7.5		0.1	6	0.2	0.1	0.1	0.2	0.2
	Liliales	Smilacaceae	71	NA		NA		NA		NA		NA	
	Poales	Cyperaceae	76	13.5	NA		NA		NA		NA		
		<i>Schoenus nigricans</i>	70	1.7	NA		3.7	22.2	8.3	8.1	2.9	16.9	15.1
	Poales	Poaceae	64	12.2	NA		21.7	NA		NA		13.1	10.5
	Apiales	Apiaceae	56	0.3	0.6		0.5	0.5	0.9	0.7	0.5	0.7	1.7
	Brassicales	Brassicaceae	52	0.1	0.1		0.1	0.1	0.1	0.2	0.1	0.1	0.3
	Dipsacales	Caprifoliaceae	72	0.3	1.5		0.6	1.2	0.4	0.8	0.4	0.2	2.2
	Ericales	Ericaceae	59	0.1	0.1		0.1	0.1	0.1	0.1	0.1	0.1	0.2
Magnoliopsida	Ericales	Ericaceae	56	4	1.1		5.5	0.9	2.9	14.5	4.6	28.9	NA
	Fabales	Fabaceae	55	0.1	0.1		0.1	0.1	0.1	0.1	0.1	0.1	0.2
	Gentianales	Rubiaceae	62	1.5	1.1		3.7	0.8	20.7	3.4	2.1	NA	
		<i>Sherardia arvensis</i>											
		<i>Cinnamomum aromaticum</i>	76	NA		NA		NA		NA		NA	
	Laurales	Lauraceae	77	NA		NA		NA		NA		NA	
		<i>Lindera obtusiloba</i>											
	Malpighiales	Euphorbiaceae	58	0.1	0.6		0.2	0.6	0.2	0.4	0.3	0.4	0.4
	Piperales	Aristolochiaceae	76	5.5	11.2		8.1	11.9	12.4	17.7	6.6	18.9	22.4
	Rosales	Rosaceae	65	0.5	0.5		1.2	0.4	1.3	1.5	0.6	0.9	3.4
Pinopsida	Sapindales	Anacardiaceae	55	0.2	0.1		0.3	0.1	0.2	0.7	0.3	0.4	1.8
	Solanales	Solanaceae	77	8.3	22.1		NA		12.1	20.3	NA		NA
		<i>Solanum citrullifolium</i>											
	Pinales	Pinaceae	59	16.4	8.4		NA		6.6	NA		NA	
		<i>Pinus strobus</i>											

Note: The lowest three values of each row have been highlighted (bold).

Appendix 7

Table A6. Required read depth for mix 2.

classification			GC [%]	required minimal read depth to achieve a detection with 95% probability (no singletons) [thousands]									
class	order	family		ITS-3p62p1F1 + ITS-4unR1	UniPlantF + UniPlantR	UniPlantF + ITS-4unR1	ITS-3p62p1F1 + UniPlantR	58SPL + ITS-4unR1	ITS-u3 + ITS-u4	ITS-p3 + ITS-p4	ITS-S2F + BEL-3	ITS3 + ITS4	
Liliopsida	Asparagales	Asparagaceae	74	NA		NA		NA		NA		NA	
	Liliales	Liliaceae	64	0.1	6.5		0.1	5	0.1	0.1	0.1	0.2	0.2
	Poales	Cyperaceae	70	2	NA		2.9	22.3	9.6	9.4	1.7	11.7	16.8
	Apiales	Apiaceae	56	0.2	0.5		0.5	0.8	0.6	0.3	0.5	1.4	
	Asterales	Asteraceae	56	0.1	0.1		0.1	0.2	0.3	0.1	0.1	0.5	
	Brassicales	Brassicaceae	52	0.1	0.1		0.1	0.1	0.2	0.1	0.1	0.3	
	Caryophyllales	Caryophyllaceae	60	0.2	0.1		0.4	0.1	1.1	0.4	0.1	0.4	0.7
Magnoliopsida	Caryophyllales	Polygonaceae	77	1	0.3		2.1	0.2	1.4	4.2	1.2	2	9.3
		<i>Calluna vulgaris</i>	62	3.3	1.2		4.4	1.2	2.2	13.9	7.1	11.4	NA
	Ericales	Ericaceae	56	3.2	0.9		5.4	0.9	3	16.9	4.5	14.6	19.8
		<i>Erica spec.</i>											
	Fabales	Fabaceae	55	0.1	0.1		0.1	0.1	0.1	0.1	0.1	0.1	0.2
		<i>Anthyllis circinnata</i>											
	Fagales	Fagaceae	68	NA		NA		NA		NA		NA	
		<i>Quercus ithaburensis</i>											
	Laurales	Lauraceae	75	19	NA		NA		NA		NA		25.8
		<i>Laurus nobilis</i>											
Pinopsida	Magnoliales	Annonaceae	79	NA		NA		NA		NA		NA	
		<i>Asimina triloba</i>											
	Malpighiales	Euphorbiaceae	58	0.1	0.5		0.2	0.5	0.1	0.4	0.2	0.2	0.4
	Rosales	Rosaceae	65	0.5	0.4		1.2	0.3	1.1	1.5	0.4	0.5	3.5
		<i>Prunus dulcis</i>											
	Solanales	Convolvulaceae	58	0.2	0.2		0.6	0.2	1.1	0.6	0.2	0.3	1.3
		<i>Convolvulus siculus</i>											

Note: The lowest three values of each row have been highlighted (bold).

Appendix 8

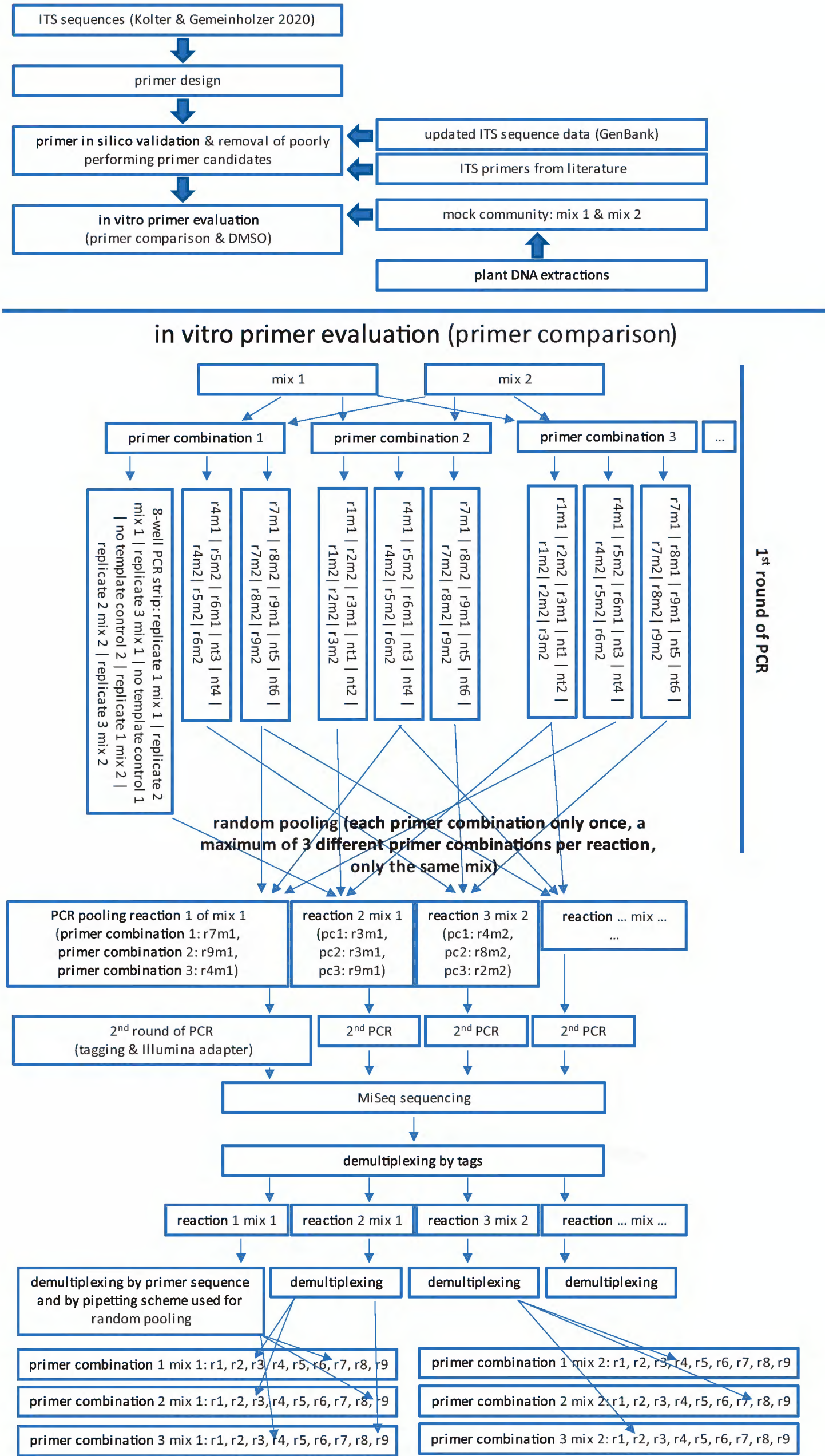


Figure A2. Experimental workflow and study design.

**Supplementary material 1****Supplementary files**

Author: Andreas Kolter, Birgit Gemeinholzer

Data type: zip. archiv

Explanation note: **Suppl. file 1.** Lists a taxonomic breakdown of the *in silico* primer mismatch testing. It furthermore contains the sequence files used for primer evaluation and the DMSO trial experiment figure. **Suppl. file 2.** Lists the result of the primer *in silico* testing and visualises them without any thresholds. **Suppl. file 3.** Contains family level alignments of the SSU, LSU and 5.8S nrDNA regions. It also contains the custom R scripts used for *in silico* testing, as well as mismatch figures on family level with a threshold of 30% in various formats. **Suppl. file 4.** Contains read number information and rarefaction curves of the mock communities. It also contains a graphical representation of the detection chance per genus per read number. **Suppl. file 5.** Contains information about the primers with ambiguities used in this study and a breakdown to all possible primer variants.

Copyright notice: This dataset is made available under the Open Database License (<http://opendatacommons.org/licenses/odbl/1.0/>). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: <https://doi.org/10.3897/mbmg.5.68155.suppl1>